FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Behavioral Analytics for Medical Decision Support: Supporting dementia diagnosis through outlier detection

João Miguel Pinto Ferreira

Mestre em Engenharia Informática e Computação

Supervisor: José Luís Borges

July 19th, 2013

Behavioral Analytics for Medical Decision Support: Supporting dementia diagnosis through outlier detection

João Miguel Pinto Ferreira

Mestre em Engenharia Informática e Computação

Approved in oral examination by:

President: Jorge Manuel Gomes Barbosa (PhD)

External Examiner: Pedro Manuel Henriques da Cunha Abreu (PhD)

Supervisor: José Luís Cabral Moura Borges (PhD)

July 19th, 2013

Abstract

We live in a world that is ageing at a rapid pace and where life expectancy is growing. This is a phenomena commonly known as demographic change and has been a focus of various fields of study. With the number of elders rapidly increasing, their problems, in particular health related ones, are a major concern. As many of our elders live alone at their homes tracking their health and behaviours is essential to diagnose and address several health problems that may arise.
Dementia is a syndrome that can be caused by several diseases being the most well known Alzheimer's Disease. It is characterized by a progressive loss of cognitive abilities that ultimately lead to a complete dependence on others to carry out even the most basic tasks of everyday living. Although there is no known cure for dementia there are some treatments that have shown interesting results in retarding its evolution. But all the results that any treatment can get are greatly dependent on an as early as possible diagnose because treatments starting at later stages have little or no effect in stopping the progress of dementia.
The first symptoms of dementia are small changes and difficulties experienced by the elder person. Those details will easily go unnoticed by the elder, family members or other caregivers and may not be duly reported to the health professional that could give a diagnose. Therefore it is important to find other ways of detecting those small changes. It is here where ubiquitous collection of data through the usage of sensors plays an important part. With a well structured sensing platform it is possible to collect detailed information. But the information itself is not enough. Therefore that data needs to be treated and analysed in order to present pertinent information, that might be valuable to the diagnosis process.
This project proposes to design a system that receives information from various different sources and analyses that data to find possible dementia signs. But before trying to find these possible signs the system establishes the normal pattern of behaviour of a person as a baseline for future analysis. In order to find those signs of dementia outlier detection techniques and algorithms are used during the analysis process. All the information that enters the system and resultant from its analysis is then stored to be available for consultation by health professionals, caregivers and elders. This information is then presented in a web visualization, that focus in showing the information in a meaningful way, without any unnecessary and distracting elements.
The system was tested using real world data sets and data produced by Java programs created specifically for each usage scenario. With the conducted tests was possible to see that this system is able to detect outliers in different types of data, after comparing the data that continuously arrived with the normal patterns established for an user. With these results it is possible to infer that such a system could be in fact a valuable tool in the diagnosis process of dementia and in monitoring its development as well.

Resumo

Vivemos num mundo cada vez mais envelhecido e onde a esperança média de vida está a crescer. Este é um fenómeno conhecido como evolução demográfica e tem sido foco de vários campos de estudo. Com o número de idosos a aumentar rapidamente, os seus problemas de saúde têm de ser considerados permentes. Muitos idosos vivem sozinhos nas suas casas e o seu acompanhamento é essencial para diagnosticar e tratar vários problemas de saúde que possam surgir.

A demência é uma síndrome que pode ser causada por várias doenças, sendo a doença de Alzheimer a causa mais comum. É caracterizada por uma perda progressiva das capacidades cognitivas que, em última análise, levam a uma completa dependência de outros para realizar as tarefas mais básicas da vida quotidiana. Embora não haja nenhuma cura conhecida para a demência existem alguns tratamentos que têm obtido resultados interessantes no retardar da sua evolução. Contudo, os resultados que possam ser obtidos com qualquer tipo de tratamento estão dependentes de um diagnóstico precoce pois tratamentos iniciados em fases mais avançadas têm pouco ou nenhum efeito em atrasar o progresso da demência.

Os primeiros sintomas de demência são pequenas mudanças e dificuldades encontradas pela pessoa idosa. Esses detalhes passam facilmente despercebidos por idosos, familiares e prestadores de cuidados não sendo devidamente comunicados ao profissional de saúde que poderia obter um diagnóstico. Portanto, é importante encontrar outras maneiras de detetar essas pequenas mudanças. É aqui que a recolha ubíqua de dados através do uso de sensores desempenha um papel importante. Recorrendo a uma plataforma de recolha de dados bem estruturada é possível obter informações detalhadas. Mas a informação em si não é suficiente pois os dados precisam de ser tratados e analisados para ser possível apresentar informação pertinente, que pode ser valiosa para o processo de diagnóstico.

Este projeto propõe desenvolver um sistema que recebe informações de várias fontes e analisa esses dados para encontrar possíveis sinais de demência. Mas antes de tentar detetar esses possíveis sinais o sistema estabelece o padrão normal de comportamento de uma pessoa como base para futuras análises. A fim de encontrar esses sinais de demência, técnicas e algoritmos de detecção de outliers são utilizados durante o processo de análise. Toda a informação que entra no sistema e resultante da sua análise é armazenada de forma a estar disponível para consulta pelos profissionais de saúde, prestadores de cuidados e pessoas idosas. Esta informação pode ser depois consultada numa visualização web, que se concentra em apresentar as informações de uma forma significativa e sem elementos desnecessários.

O sistema foi testado utilizando conjuntos de dados do mundo real e dados produzidos por programas Java criados especificamente para cada cenário de uso. Com os testes realizados foi possível perceber que este sistema foi capaz de detetar informações anormais em diferentes tipos de dados, depois de comparar os dados produzidos de forma contínua com os padrões normais estabelecidos para um utilizador. Com estes resultados é possível inferir que este sistema poderia ser de facto uma ferramenta valiosa no processo de diagnóstico de demência e na monitorização do seu desenvolvimento.

Acknowledgements

I would like to thank my two supervisors, José Luis Borges from FEUP and Klaus Schaefers from Fraunhofer, for their continuous guidance and support through the design, development and writing involved in this dissertation work. Furthermore I would like to thank them for continuously challenge me to improve my work in all its fronts by constructively criticising its weakest parts. Also I have to thank my English teacher, Ana Veiga, that spent the time reading my work, that sometimes might become boring for her, as she is not from the IT area.

I also have to thank all my co-workers at Fraunhofer. I was well received by all that already worked there and made my path alongside the ones that entered at the same time as me, to also work in their dissertations. To them I would like to thank for being always ready to help when I needed.

To my friends I would like to thank all the good times we had together, all the fun, all the nights, all the dinners that allowed me to clear my mind and relax a little from all my work.

A very special thanks to my girlfriend for always believing in me, sometimes more than I believed myself. It was her support that kept me always moving forward with the full belief that my goals were all reachable. Furthermore I have to thank her for helping me relax when I could not do it by myself in spite of sorely needing it.

Finally would also like to thank all my family. In particular my mother and my sister my for always putting up with me during the work, even when I was not in my best mood. And last but not least to my father, who passed away last year, I have to thank all the guidance and lessons taught while he was by my side, they made me who I am. I hope I am making you proud.

“That’s the trouble with you young people.
You think because you ain’t been here long, you know everything.
In my life I already forgot more than you ever know.”
Neil Gaiman, Anansi Boys

Contents

List of Figures

List of Tables

LIST OF TABLES

Abbreviations

AD	Alzheimer's Disease
AAL	Ambient Assisted Living
VaD	Vascular Dementia
DLB	Dementia with Lewy Bodies
FTD	Frontotemporal Dementia
IADLs	Instrumental Activities of Daily Living
GPS	Global Positioning System
GPRS	General Packet Radio Service
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
GSP	Generalized Sequential Pattern
REST	Representational State Transfer
JSON	JavaScript Object Notation
URL	Uniform Resource Locator
NoSQL	Not only SQL
SQL	Structured Query Language

Chapter 1

Introduction

We live in a continuously ageing society, where the most senior age group (aged 85 or older) has shown the most growth of all age segments. Since 1950 the probability of surviving from age 80 to 90 years old has risen from 16% for women and 12% for men to 37% and 25% respectively in 2002. And it is clear that we have not reached a limit, being forecast a further rise in life expectancy, as we can see in Figure 1.1. Most new born in developed countries since the year 2000 will reach the 100 years mark, and so it becomes obvious that addressing the issues, especially the health related ones, involving this population is a problem at hand. [CDRV09] All these results in the phenomena known as demographic change that has been studied with more focus in the past few years.

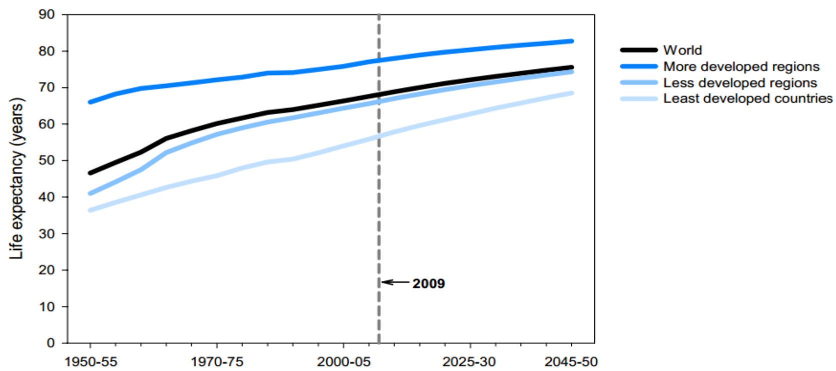


Figure 1.1: Evolution of life expectancy

The correlation between age and dementia is well established, being age one of the major risk factors. And having the world population over 65 doubling from 35 million today to 70 million by 2030 makes understanding dementia, its social and economic costs, one of the challenges, not only for the future, but especially for the present. In one way or another we will all experience dementia by suffering from it ourselves or dealing with it, with a family member, being a caregiver or having a social or professional encounter with an elderly person.

There is also the monetary cost associated with dementia. In 2007 it was estimated that $315 billion was being spent each year. [Sha09] That value quickly escalated to $604 billion in 2010,

being only 16% of those related with direct medical costs. [OI12, p.8]
Ideally any diagnose system should be as seamless as possible allowing the person to carry on it's normal day life with no obstruction from technology in order to not introduce biases in the results. There are nowadays several examples of systems that collect information and aid us in everyday life, even though we do not always understand the complexity behind it. This is how a dementia diagnose tool should work, having complex data mining and computation, but seeming invisible to the patient.
Ambient Assisted Living (AAL) solutions offer the possibility of collecting all kinds of data from home installed sensors, that are unobtrusive to the user and give the opportunity of having a different insight to the person everyday. Therefore using concepts from AAL is a starting point to develop a system that purposes itself to be a diagnosis tool. In addition to a home environment capable of collecting information of a person it is also important the ability to collect some information about the everyday life of a person outside his home. Here come into play other technological aids like wearable sensors and especially smartphones that are nowadays equipped with all kinds of sensors that can record a lot of useful information without causing any kind of distress to the person who carries it.

1.1 Problem Description and Motivation

Having established that our world is indeed ageing, and that the correlation between age and dementia is clear, one consequence comes to mind - The number of people affected by dementia will rise in the future. And if a large amount of money is already being spent worldwide, those costs will tend to increase.
It becomes clear the importance of an early diagnose and here is where technology can have a huge part to play. Even more if it is taken into consideration that an early diagnose is the best way to retard this syndrome, that is incurable, and that figuring out those first symptoms can be a challenge for the family, healthcare professionals and even to the person themselves. [Naz11]
Furthermore it is important to keep in mind that many of our elders, specially in big cities, live alone at their homes, a fact that can make the diagnosis an even harder process.
There has been a growing concern with the elders living by themselves, as their numbers rise. There are groups that visit a person at home to check on them regularly, phone support lines and day centres that are focused on accompanying older people that live on their homes alone. Some of them even provide meals to those people in order to avoid, as much as possible, forcing them to cook, as this can be a burdensome, demanding and even dangerous task. This type of assist is invaluable to those people but is limited in the sense that consists in only a period during the day, not allowing to monitor and help the elder through the whole day. Ubiquitous and unobtrusive sensing can offer an opportunity of accompanying a person in a continuous manner, even from a distance. Collecting data from a myriad of sensors offers information that otherwise would not be available and also allows to take a retrospective to some aspects that could be overlooked.
There are different types of data that may be useful to monitor in order to help in the process of

diagnosing diseases, in particular dementia related ones that manifest themselves with behavioural symptoms. For example tracking the outside movement of a person with dementia may show signs of disorientation. Inside the home environment it may be possible to sense the way a person goes about his normal day routine. And by monitoring patterns of movement may be possible to detect signs of unwanted behaviours such as night wandering. Furthermore registering the number of times a person leaves home may allow to perceive signs of social withdrawal. This are some common early signs of dementia that may become visible when the elder is continuously monitored. If an elder being monitored shows any of these signs than this information may help the health professional to reach a diagnosis of a type of dementia.

It is of vital importance that the technology does not get in the way of the everyday life of a person. Most elders are not very comfortable dealing with computers, smartphones and other technological tools so their interaction with them must be as simple as possible and limited to a minimum. In a monitoring tool the interaction can be reduced to the consult of the website to check the collected data. This access must be as simple as possible and the data should be well organized and well perceived. All the other components can, and should be hidden from the user in order to not cause any distress.

Furthermore, this simplicity in accessing information should also be extended to family members, caregivers and medical doctors. It is important that these groups of persons have access to the information, obviously when allowed by the elder in case he is capable of making this decision. By accessing the information these people can provide better care and also keep a track record of possible changes of behaviour that may be seen as cause for concern. With better information doctors will be able to make more accurate diagnosis and caregivers and family members may gain some peace of mind by having further information about the elder.

In spite of all the benefits of continuous monitoring the elders behaviours this imposes some ethical problems, that have been duly discussed by several authors. The most obvious and serious problem is the restrain of a person freedom. If an elder has its activities continuously monitored then it may be considered that his freedom is being clearly jeopardized, and, as the right of freedom is one of the most basic of human rights, this issue needs to be considered carefully. [WHOD03] Despite this there have been studies that show that not only the caregivers but also the care receivers feel more comfortable, secure and even more free when they are being monitored. For instance in the work of Pot et. al [PWH12] caregivers said that they tended to allow the persons with dementia to go outside alone more often (60%) when they were using the tracking device, therefore giving them a little more freedom, even if a controlled one.

1.2 Objectives

Dementia is degenerative and progressive with its symptoms starting as subtle behavioural changes and mild impairment and progressively evolving into serious incapacitation and in the final stages total dependence of a caregiver to go through the everyday living. So, in order to get an early and accurate diagnose, it is important to establish the person's base behavioural pattern and use it to compare with future patterns in order look for the changes that might be interpreted as signs of dementia. With this it is possible for the health professional to achieve a more accurate diagnose because he can understand if a possible sign, such as social withdrawal, is a real sign of dementia or only a personality feature. [FLGGV+11] And as dementia is progressive, even if in the base pattern there are already some warning signs those will tend to deteriorate and will be detected as a change in the person's behavioural pattern. AAL solutions may offer the possibility of being a part of the solution collecting data that after treated and analysed may be used as a medical decision support tool. [MAMM12]

Therefore to find these behavioural patterns is essential to collect data from various sources such as smart home sensors, wearable sensors, GPS tracking devices and others. But more than collecting data it is important to give meaning to that data by analysing it and uncover those normal patterns for a user. Furthermore, having the normal pattern established will then be possible to find abnormal data that may appear, data that does not comply with the normal pattern. This abnormal data, that may be possibly related with dementia signs, will be detected using outlier detection techniques.

The objective of this project was to develop a system capable of receiving and analysing data gathered by several sensors and then presenting it, and the results obtained, to all the users. The users will be the elders being monitored, caregivers and also health professionals. The system aims to aid the health professionals in achieving an accurate and as soon as possible diagnose by giving them detailed information and highlighting possible signs of dementia. Adding to be an aid to the diagnosis process the system also aims to help monitoring and tracking the evolution of the condition in elders that may already be diagnosed.

The system envisioned will need to handle several streams of data continuously being produced by its users. This data needs to be received by the system and forwarded to the right analysis process. The analysis process needs to be done in real time in order to have all the results available to be consulted. To achieve a real time analysis process this was done in a distributed fashion, implementing some outlier detection algorithms and techniques in a distributed computation system that takes advantage of the use of clusters of machines. All the data received and resultant from the analysis needs to be stored. To handle the large number of write operations that will need to be done, as we are dealing with continuous streams of data, the database used also works in a distributed way. Finally all the information needs to be presented to the users. To do this all the information stored can be presented in a visualization that focus in giving all the information valuable to the user with no unnecessary information that might make it harder to understand the information.

In summary this project aims to develop a platform capable of:

- Handle tracking information of the user outside movement collected with GPS location;
- Handle home behaviour data collected through sensors;
- Respect ubiquitous computing principles (be seamless);

- Create behavioural patterns with the mined data;
- Detect outliers in behavioural patterns;
- Understand if those outliers may be interpreted as dementia signs;
- Present information collected and analysed in a simple and meaningful way

Being the main goal of the whole project to develop a system capable of analysing different types of data and being an aid to obtain an early and accurate diagnose as well as a good monitoring of evolution of the condition, by giving detailed and analysed information to its users.

1.3 Document structure

Besides this introductory chapter this dissertation is composed by six more chapters. In Chapter 2 it is presented some detailed information about dementia, its different types and their singularities, the evolution of the number of sufferers, how it affects elders and other third parties and what is the importance of an early diagnose. Chapter 3 contains a list of projects that may be considered the state of the art in monitoring the health condition of elders, particularly the ones who suffer from cognitive impairments. In addition it is presented a table comparing all the referred projects and contextualizing the solution proposed with the products already existent. In Chapter 4 are presented several methods and algorithms that may be used in outlier detection, this is finding abnormal points on the data gathered. Chapter 5 presents all the information about the development of this project, starting with the specification and the system architecture designed and then giving some details about the implementation of the system specified. In Chapter 6 are presented the tests made to validate the system by doing analysis of different types of data and discussing the results obtained. Finally Chapter 7 presents the conclusions from all the work made, referring the contribution of this project to the problem at hand and explaining future work that would be valuable to improve the project.

Chapter 2

Dementia and the elderly

Dementia is not a disease but actually a syndrome related with a series of symptoms caused by several types of brain diseases, such as Alzheimer's Disease (AD). Indeed it is estimated that AD accounts for 60%-70% of all cases of dementia, with the other most common causes being Vascular Dementia (VaD), Dementia with Lewys Bodies (DLB) and Frontotemporal dementia(FTD). [OI12, p.7] But identifying the underline disease that is the cause of dementia is not straightforward because most cases are mixed ones. [OI12, p.19,20] The most exposed to dementia are the elderly, 14% of all people over 71 suffer from some form of dementia [Sha09] and being cases of persons suffering from dementia under 65 rare and considered early onset cases. Early onset dementia is considered uncommon and are usually related with genetic cause. With a world rapidly ageing in the past years, and showing signs that this is a trend to continue, especially for countries still in development, [CDRV09] the numbers of people that suffer from dementia will also tend to rise. In 2010 was estimated that 7.7 million cases were diagnosed each year and 35.6 million people were living with dementia. Furthermore, it is expected that this number will double each 20 years and by 2050 we may have reached the staggering number of 115.4 million people living with dementia. [OI12, p.4,p.12] [FLGGV$^+$11] Presented in Figure 2.1 is a graphic that shows the projected evolution of the numbers of people with dementia both in high and low income countries.

Dementia is degenerative, progressive, incurable and unstoppable. However it's worse effects can be retarded, and there are a number of clinical trials testing new therapies that hold promise of helping in further retarding the progress of dementia. [Int09] There are a wide range of symptoms associated with dementia that can be perceived in a number of changes such as: [Naz11]

- Congitive changes

  Seen as disorientation, both spacial and temporal, loss of short-term memory, difficulty in language skills, written and spoken and others;

- Day-to-day skills changes

  Loss of ability to complete common day-to-day activities ranging from driving to cooking or shopping;

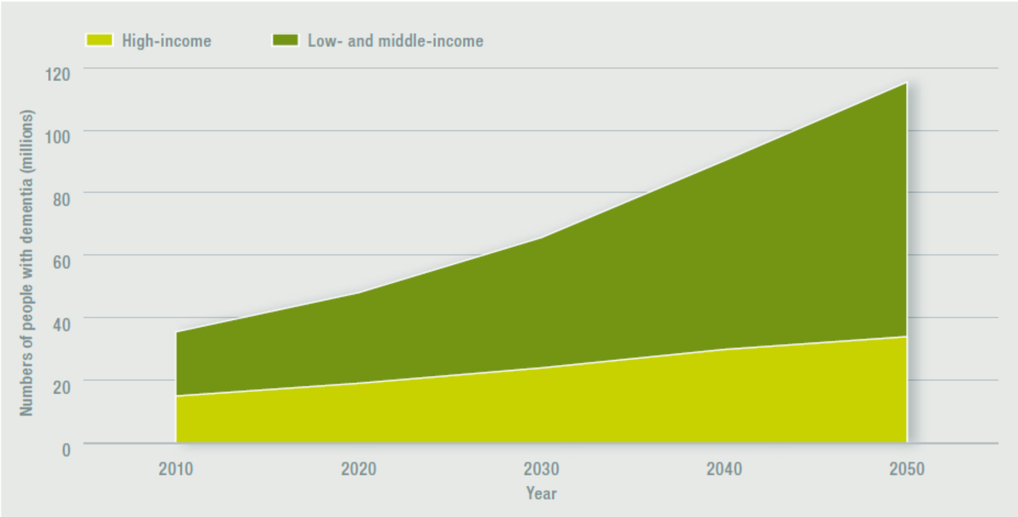


Figure 2.1: Evolution of dementia patients in low and high income coutries [OI12]

- Personality changes

  Some personality features may be changed creating a number of consequences, some of them contrary to each other, such as social withdrawal, disinterest, frustration, excessive flirtatiousness and inappropriate friendliness;

- Problematic behaviour

  Appearance of some abnormal behaviour like agitation, restlessness, wandering and being out of bed at night;

- Psychiatric symptons

  A myriad of symptoms such as hallucinations, paranoia, depression, anxiety and fearfulness;

Even though some of these symptoms, like paranoia or total disorientation, seem suitable for a straightforward diagnose, they are related with later stages of the disease, and earlier ones, for instance forgetfulness, are easily dismissed as part of normal ageing process.

2.1 The different types of dementia

There are several different types of dementia each one with different causes, different ways of treatment and different symptoms, even if many of the symptoms are present in all the types. As mentioned above the most common type of dementia is AD. But it is important to understand that most AD cases are, sooner or later, proved not to be pure, but in fact are mixed with some other type of dementia, in most cases VaD. And it is not only AD but all the most common causes of dementia, referred above, that present themselves as mixed kinds. [OI12, p.19,20]
Even taken into consideration that most cases of dementia are mixed ones, it is important to understand the differences between AD, VaD, DLB and FTD, the most prominent forms, both in terms

of causes and symptoms. In the list below are presented the most common types of dementia and their symptoms:

- Alzheimer's Disease (AD)

  AD is caused by neurofibrillary tangles and cortical amyloid plaques. Those tangles and plaques damage nerve cells and affect the transmission of messages trough the brain, resulting in learning and memory impairments. The appearance of the disease is commonly related with genetic causes that are responsible for the presence of the APOE*E4 gene that is related with a high number of amyloid plaques that ultimately causes the disruption of brain function. The most common symptoms associated with AD are loss of short term memory, disorientation, decline in decision making capabilities, apathy and depression. [Sha09, Int09]

- Vascular Dementia (VaD)

  Unlike the other forms of dementia, VaD can be either progressive or acute. This depends if it is caused by a sudden infarction of a large blood vessel or by a series of smaller infarctions. The symptoms related with VaD are similar to AD's with more emphasis in the mood fluctuations and less memory impairments. Adding to those you can include body weakness, rapid small steps and speech difficulties. [Sha09, Int09, Naz11]

- Dementia with Lewy Bodies (DLB)

  Lewy bodies are small deposits of protein in the nerve cells that disrupt messages that go through our brain and are also found in people with Parkinson's Disease. In fact DLB includes Parkinson's Disease Dementia. Adding to the common symptoms associated with Parkinson's Disease, like tremors, rigidity and loss of balance, persons with DLB frequently have hallucinations and present severe fluctuations in cognitive ability, attention and concentration. [Sha09, Int09, Naz11]

- FrontoTemporal dementia (FTD)

  FTD is related to damages to the frontal and temporal lobes of the brain that may derive for a number of pathologies. FTD has a unique set of behavioural symptoms like disinhibition, apathy, mood changes, repetitive behaviour and changes in eating habits. Further symptoms are speech difficulties and changes in appetite. [Hol12, Int09, MC05]

More than just identifying the presence of dementia is important to have as much information as possible to understand what disease is causing it. Some symptoms are transversal to all of the most common causes of dementia but others may help in accurately identifying the type of dementia we are dealing with, so having as much information as possible will always be an important factor in differentiating between them.

2.2 Importance of an early diagnose

As in any other disease or health problem, a diagnose is essential to begin addressing the issues related with dementia, that affect the patients, their families and caregivers. But this concern goes even deeper with dementia, a severely under-diagnosed issue, where 28 of the 36 million people affected are not diagnosed which creates a treatment gap that needs to be closed [Int11, p.7,10]. Furthermore, because of the progressive nature of this syndrome [Sha09], some benefits in an early diagnose may be found.

Before exploring the possibilities given by an early diagnose it is important to understand what is considered an early diagnose. Nowadays, apart from all the cases that are not diagnosed, most diagnoses are made at a late stage where behavioural symptoms and cognitive impairment are well marked, severe disability to the person with dementia is visible and great amount of stress affect the families that need to cope with them. Therefore it is easy to understand that nowadays most cases don't even have a timely diagnose. A timely diagnose would be one that responded to the concerns of older people and their families. So an early diagnose may be considered one that is made before this stage, and that is already made possible by technology and healthcare structures, but rarely occurs. This happens because there is still a long way to go in raising awareness towards dementia and with that promote more help-seeking by persons with dementia and their families.[Int11, p. 11-13,14]

Now that the concept of early diagnose has been analysed, its possible benefits need to be understood. First to refer that there is little quantitative data supporting the benefits of an early diagnose. Despite this, there is the general expert opinion that an early diagnose is clearly beneficial and should be promoted. The most obvious benefit of an early diagnose is the access to expert medical help and treatment in a timely manner that can retard the most serious effects of dementia with better clinical outcome. Another very important benefit that can come from a diagnose made in an early stage is the opportunity that is given to persons with dementia and their families to adapt and adjust to the reality they will have to face. Furthermore the relief that comes from the better understanding of the symptoms reduces the anxiety that they generally cause. It also gives the opportunity for the person to get their affairs in order, such as power of attorney, financial situation and even decisions for the course of treatment to take. An early diagnose also permits an effective risk reduction in a myriad of situations such as driving, medication errors or forgetfulness and even accidents at home. Finally the cost reduction that can come from the early intervention by delaying the time when the person will have to be committed to assisted care, thanks to the better results that outcome from treatment.[Int11, p. 24-30]

Adding to this is also important to take into consideration that there are many different types of diseases behind a dementia diagnose. Accurately identifying what is the cause may prove itself a tricky question as some symptoms are very similar in different types of dementia. But having a correct diagnosis, in addition to an early one, is essential to further improve the results associated with any treatment that may help retard the development of the disease.

Chapter 3

Related Work

Ambient Assisted Living (AAL) solutions offer a wide a range of opportunities to collect behavioural data to study a myriad of subjects. Furthermore, the diagnosis and understanding of diseases and health problems has progressively invested in using unobtrusive sensing of a person's behaviour to collect data that is otherwise difficult to collect. Smarthome environments, wearable sensors and using other common objects, such as smartphones, as pervasive sensing platforms has created an opportunity to collect large amounts of data that can offer important new insights. With this data made available the existence of tools that help to organize, understand and interpret it becomes essential. But more than collecting data there has been an investment in creating AAL facilities that help improving quality of life for persons of all ages, with more focus on the elder and in persons with cognitive and physical difficulties. These facilities can even offer the possibility to adapt to changes in behaviour of a person, conveying messages and other forms of positive reinforcement to make people avoid negative behaviours.
There has also been a clear effort in the past years to develop a number of applications to collect data in order to promote better quality of life to people of all ages. This started with applications where people manually logged all kinds of data from food intake to hours of sleep. These applications are widely available nowadays at application stores like Google Play or Apple's App Store. Doing this is commonly described by the term of *Quantified Self* and has people that follow this religiously and even meet in conferences to share their experiences. [Lab] The problem with this manual logging is that it can easily become burdensome to the user and biases can be easily introduced by mistakes or forgetfulness. This becomes even more relevant if you think about elder people, especially the ones who suffer from increasingly greater loss of cognitive functions, strongly related with dementia. So here comes into play the power of pervasive and ubiquitous computing. Using continuous monitoring, through the use of sensors, we can alleviate the user that does not need to log the information since this is being done ubiquitously by the system itself. And when we take into consideration that traditional methods of diagnosis of dementia, still the most used nowadays, usually depend on either self-reports, caregiver reports or performance testing by occupational therapist it is easy to see an opportunity to unobtrusive sensing in being a source of continuous. This fine grained information may prove itself valuable in the diagnosis pro-

cess. [Lee10] There have been a number of works focusing on detecting dementia. Some focus on detecting how well everyday living tasks are completed [Lee10, ANMF10], others in tracking outside movement to detect abnormal patterns, wandering and getting lost [SAF+08, Mis05, PWH12] and finally some collect behavioural data and even recognise and interpret behavioral patterns. [FLGGV+11]
Some other work, not dementia related, has been made but that uses different types of sensor like GPS, accelerometer, and others, that recognize activities and create summaries to track and evaluate the well being of a person.

3.1 State of the art

In this section the projects that can be considered as reference in tracking and monitoring illnesses, in particular, but not limited to, dementia related ones will be discussed. All these projects offer the possibility of having a different insight to the habits and behaviours of the persons who use it. This new view can be invaluable to several areas of concern in cases of illnesses that have behavioural symptoms, like dementia. First, in terms of diagnosis, having access to the great amount of data originated by continuous monitoring can provide information that otherwise would not be available to health professionals. To family members and caregivers having a mean to check on the elder behaviours can result in a greater sense of security and peace of mind. Finally to the patients, and particularly to elder ones, this kind of information can help them to better accept their condition and find ways of coping with it.
These projects rely on home sensors, smartphone sensors, besides exploring other smartphone capabilities, GPS tracking and sensors built in to everyday objects to ubiquitously collect and organize the data in a way that may be perceived by a normal person.

3.1.1 eMotiva

This is a project being developed in Spain implemented in nursing homes that takes advantage of Ambient Assisted Living facilities to monitor the elder behaviour, establish a base behavioural pattern. From this point on the system tries to detect changes in that pattern that may be considered signs of dementia. After detecting this erroneous behaviour the system informs the Health professional that has access to a workflow chart representing the patterns and what has changed in it. With this information the Ambient Assisted Living component may be put in practice in order to improve motivation and quality of life of the elder. This motivation is decided by the Health Professionals, but they can define motivating action to be performed automatically whenever a certain deviation from the pattern is detect.
All the data collected to create the workflow charts and determine the behavioural pattern is ubiquitously and unobtrusively gathered through a number of sensors in the facilities. This gives the elder all the benefits of technology without having to be in contact with it, which may be confusing to them. [FLGGV+11]
In Figure 3.1 is presented an example of the workflow charts that represent the behaviour of an

elder. In green are the behaviours that do not present signs of decline and in red the behaviours that can be considered cause for concern.

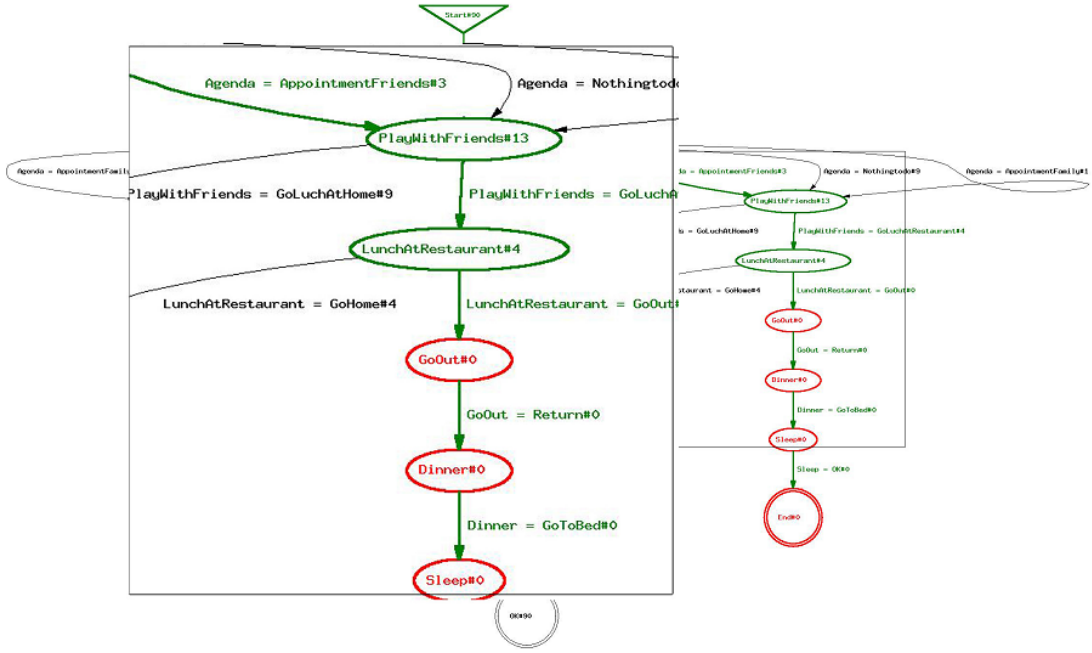


Figure 3.1: Example of eMotiva Workflow Chart [FLGGV$^+$11]

Further information about eMotiva can be found at:
`http://www.proyectoemotiva.org/`

3.1.2 GINGER.io

GINGER.io is a tracking and monitoring platform that takes advantage of the powerful sensing capabilities offered by most smartphones nowadays. This tool is divided in two different parts, a smartphone application that gathers data from several sensors, as seen in Figure 3.2b and a web based dashboard where health professionals and researchers can access the data collected, like the one in Figure 3.2a. The smartphone application continuously and ubiquitously collects data from smartphone available sensors, in this way creating a great amount of data without being a burden to the patient. Furthermore, it also allows also active data collection through user reports. All the behavioural data collected is analysed and presented through the usage of Google Analytics tool in order to ease the process of analysing the data for health professionals and researchers.

Despite not being focused only in dementia GINGER.io clearly offers the tools to help track and diagnose dementia patients. And this was not missed by their team, as they were finalists and ultimately winners of the Janssen's Alzheimer's Challenge 2012 that was seeking for new tools to improve Alzheimer's care. With their behavioural analytics platform they offered an unobtrusive, pervasive way of tracking Alzheimer's patients and even provide insight to improve quality of diagnosis.

More information about GINGER.io can be found at:

(a) Web based dashboard (b) Android application

Figure 3.2: GINGER.io interfaces

`http://ginger.io`

More information about Janssen's Alzheimer's Challenge 2012 at:

`https://www.alzheimerschallenge2012.com/`

3.1.3 Embedded Assessment of Wellness with Smart Home Sensors (2010)

In this work [Lee10] focus is given to the analysis of the changes on the ability of a person to carry out tasks that are instrumental to the everyday living. These tasks are commonly known as Instrumental Activities of Daily Living (IADLs) and can be making a pot of coffee, dial a phone number or take the medication. This work has the objective of collecting data from, not only the completion of IADLs, but also how well they are completed. For example, a person can be still taking their medication correctly but taking more time in choosing which pills to take, and this may be an indicator of onset cognitive problems. This task performance analysis is achieved by tracking the individual steps that need to be carried out to complete the task. As the data collected by the sensors can be overwhelming in quantity, salient summaries of the activities are created. Occupational therapists found the finer grained data collected useful to identify particular deficits and doctors welcomed the more detailed information than the one available from patient or caregiver interview.

3.1.4 The use of advanced tracking technologies for the analysis of mobility in Alzheimer's disease and related cognitive diseases(2008)

This project [SAF+08] proposes to use GPS to track outside movement in order to detect patterns that may be perceived as possible dementia signs, such as wandering. Out of home mobility problems is one of the most common behavioural manifestations of dementia related disorders. 360 persons over 65 years, comprising demented, mildly cognitively impaired and non impaired elders were recruited from various sources. The research will span for a period of 5 years.

The data collection is carried out in 3 waves, one year apart from each other, starting in year two after collecting all the necessary information and authorizations from the participants in the first year. After having the information comparisons will be made between the data collected in each wave. Special attention will be given to the data of the elders that presented mild cognitive impairment in the beginning of the project because by wave 3 it is expected that most of them will present a more serious form of cognitive impairment and therefore a different out of home mobility pattern. To note that this project is still in its course so conclusions are not yet available.

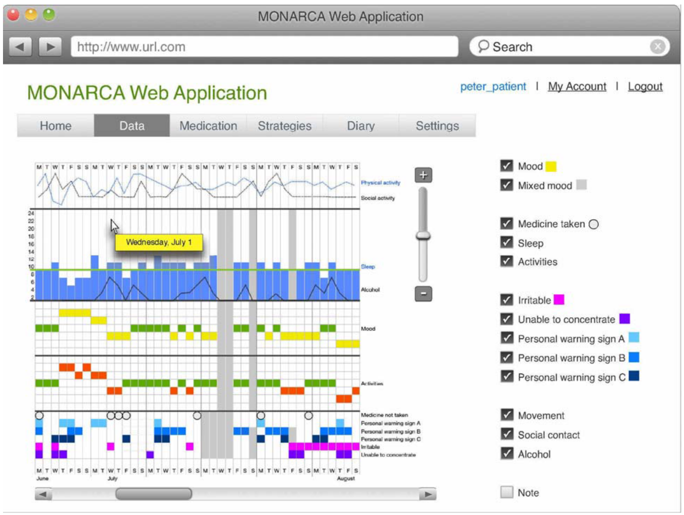
3.1.5 Electronic Tracking of Patients with Dementia and Wandering Using Mobile Phone Technology (2005)

In this work [Mis05] 11 participants were given a mobile phone to use as a tracking device. A caregiver or relative was trained to set up the phone and make sure that the patient was using it correctly.
In the first steps the reliability of the GPS tracking was studied and was decided that the optimal position to carry the phone was a shoulder bag. To further test the performance random walk simulations were conducted with volunteer staff carrying the phone. After this test it was estimated that the position information was available 87% of the time, except when in indoors or in public transportation.
Participants trials were then conducted. Interviews with the relatives or caretakers were made in order to establish the patient daily/weekly activities. Comparing this information with the data recorded it was established that they were in good agreement, what allowed to conclude that technically the tracking system worked well. To note that during the trials five participants quit due to usability or comfort issues. Important to note that 9 of the 11 participants complained about the size and weight of the device and 27% said that it was too physically demanding to carry.

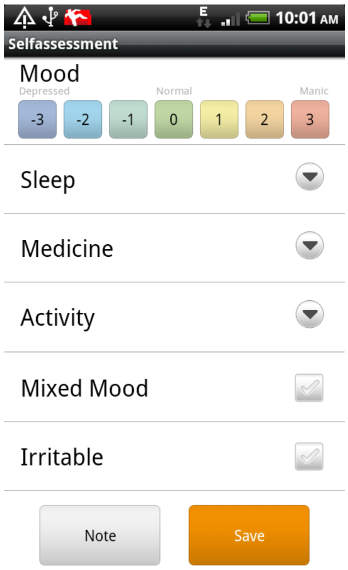
3.1.6 MONARCA

MONARCA [BFSM12] is a smartphone application that proposes to monitor Bipolar Disease patients both passive, through pervasive sensing, and actively, by user input. Data from those inputs can be viewed in the web platform by patient and clinician. Examples of the interfaces of the smartphone application, seen on Figure 3.3b, and web platform, Figure 3.3a can be seen in Figure 3.3.
The whole system has some key features. First the self-assessment that consists in answering less than 10 items about mood, sleep, level of activity and medication, reminded by the system when needed. There is also activity monitoring that is possible through the data collected from sensors available in the smartphone and from call and message history. It is also possible to have an historical view of the data collected, choosing to restraint that data to a date, or to a group of dates that may be considered relevant. This can be done by both patient and clinician. There are triggers that are actioned to notify both patient and clinician when warning signals appear. Corrective measures are given to the patient by the clinician to correct the warning signs detected. The patient can choose to share the data with family members and caregivers in order to have better

support from them. It is possible that the user will choose not to share the data and is important to have this option in order to respect the user privacy rights.



(a) Web display of patient data



(b) Android application

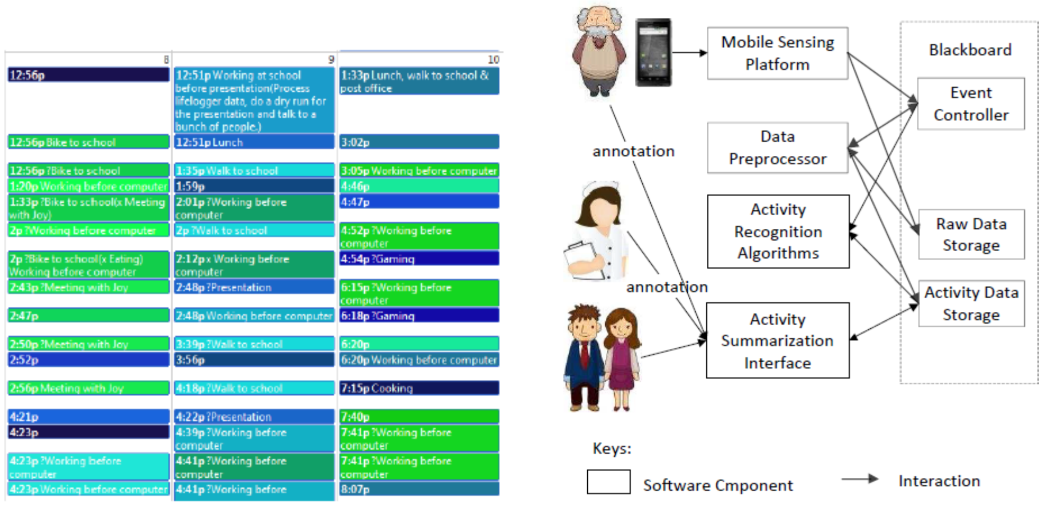
Figure 3.3: MONARCA interfaces

A small video about MONARCA:
`http://www.youtube.com/watch?v=UVS0cAxlQxM`

3.1.7 SensCare: Semi-Automatic Activity Summarization System for Elderly Care (2012)

Senscare [WPZZ12] is an Android smartphone application that unobtrusively collects data from various sensors in order to detect and identify the several activities carried out by the user during the day.
To achieve the goal of activity recognition the data from different sensors needs to be combined, because, using only data from a single source, two distinct activities can appear to be the same. One example given is eating and driving. Using only data from the accelerometer they are identical but adding GPS data of positioning and speed they become clearly different from one another. As the person carries the phone the whole day, it will serve as a sensing platform, taking advantage of the several sensors available, to record data that is sent to the cloud where it will be interpreted and divided in the activities performed by the user. All information collected by the system can be viewed by the user as a timeline or hierarchical view of the activities performed during the day. In Figure 3.4a can be found an example of the timeline view. The smartphone was chosen as the sensing platform because they are accessible to almost everyone nowadays and they offer a large number of sensors and possibilities. Senscare uses the accelerometer, magnetometer, GPS, microphone, temperature sensor to collect all the data and make the needed inferences to establish the activities being carried out. It also uses WiFi signal strength to infer indoor positioning. In

Figure 3.4b can be found a basic overview of Senscare architecture.
A series of tests were conducted to evaluate the precision and effectiveness of the system as well as feasibility of different phone carrying positions, arm or pocket. Data was collected through the course of 5 consecutive days where the phone was carried by two graduate students, that were asked to refrain from vigorous physical activity in order to simulate the lifestyle of a older person, the target group age of this project. With this they concluded that the activities recognised by the phone corresponded, most of the times, to the activities really being performed by the user. It was also concluded that the arm position was the most accurate for the majority of activities except for some specific events like walking or cycling.



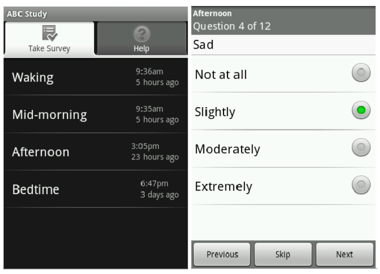
(a) Web timeline activity display (b) System Architecture

Figure 3.4: Senscare Interface and Architecture

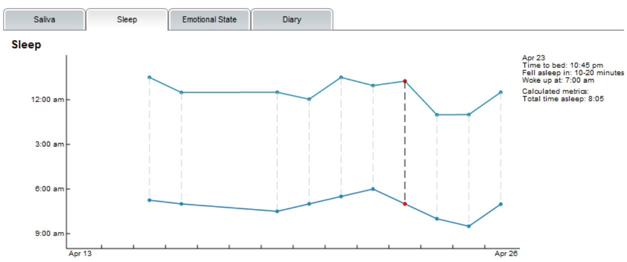
3.1.8 AndWellness: An Open Mobile System for Activity and Experience Sampling (2010)

Andwellness [HRK+10] is a system composed by an Android application, a server and a web based dashboard. The application collects data both ubiquitously, through smartphone sensors, and actively, using questionnaires prompted to the user at appropriate times that can be defined by the user in the web interface. This is important in terms of the usability because users normally do not want to be disturbed at inconvenient times or more often than necessary. Furthermore, the application was design to not interfere with the phone normal functions such as calls, other running applications or battery life.
The data collection module is composed by three components, the campaigns, the sensors and the triggers. A campaign is a study where the researcher can define the data types to be collected and surveys to be prompted to the user. In a campaign there are different roles that give different type of access to the data. The administrator that can adjust the parameter definitions of the study, researcher that can see the data from all the participants in the study and the participant that can

see its own data. The sensors collect continuous data to be analysed in order to make available a complete analysis of the data obtained from the surveys. The sensors include the ones available in the phone and others defined by the researcher. This other sensors can wireless transfer data to the cellphone via bluetooth. Finally the triggers send a notification to the user when he is supposed to answer a survey. These triggers can be defined to go off at a specific time or location by the researcher, and can even be defined individually to each user.
The dashboard offers two different functionalities. First is in the dashboard that the researchers can configure and adjust the parameters of their study, and participants can define the best time to answer their surveys, if possible in that particular campaign. The other functionality is the data collected visualization. Currently it is possible to have a graphical view of the data collected from the sensors and from the surveys as well. This works as a motivation to participants, as they can be aware of their data and improve behaviours they consider negative. To the researchers it is useful to easily obtain the data collected from all patients and take some quick conclusions. Although not yet implemented is planed to offer the users a map view of the data collected from the GPS.
Currently Andwellness has collected feedback from some sources. From Focus Groups several conclusions were drawn. First it became clear the importance of customization, especially the times of the surveys. Privacy is a key concern and users want to be able to choose with who they share their information. The importance of using non textual motivators, like images, that helps to improve user behaviours. Users also showed interest in sharing and viewing anonymized and aggregated information, in order to improve themselves with input from others. Finally it was mentioned that in order to stay motivated it is essential to the participants to feel like they are part of something useful and important. Further feedback is expected from studies where Andwellness is already deployed one involving cancer survivors and the other about risky behaviours and HIV transmission.
Figure 3.5a is a view of the surveys that can be prompted to the user and in Figure 3.5b it is possible to see the data collected represented in the web dashboard.



(a) Survey taken in Android application



(b) Web dashboard information view

Figure 3.5: Andwellness user interfaces

3.1.9 eCAALYX: Enhanced Complete Ambient Assisted Living Experiment

The eCAALYX project is an European funded project under the Ambient Assisted Living Joint Programme and is built on top of the previously existing CAALYX project, also partially funded by the European Commission, taking advantage of the infrastructures and functionalities already implemented. [BLA+09]
In the CAALYX project [BRM+07] it was aimed to develop a system capable of unobtrusively monitor the everyday of an elderly person and warn the caretaker when some possible cause of alarm occurs, also sending the location of the elder. More specifically the objectives of CAALYX were:

- Identify which vital signs and patterns were essential to determine critical states;
- Create an electronic device to monitor vital signs and detect falls. This device should also know the location of the elder in order to send this information to the caretaker when some situation happens;
- Help caretakers monitor groups of older people assigned to them and decide if a warning is reason to make an emergency call.

To achieve these goals three main areas of development were considered. First the Roaming Monitoring System that was responsible to monitor everyday activities both inside and outside the house and also know the location of the elder at all times. The Home Monitoring System, intended to further monitor activities performed inside the house, through the use of sensors and other monitoring devices. Furthermore it also creates some home automation in order to facilitate the activities of the elder and allow more social interaction through video conferences with friends and family. Finally the The Central Care Service and Monitoring System responsible to receive the alerts and give the information to the caretakers in order to let them make an informed decision if a emergency call is needed or not.
As aforementioned eCAALYX had as a start point the CAALYX project aiming to further develop it. Just like CAALYX the objective is to help monitor and improve quality of life of older persons. The main goals of eCAALYX were the following:

- Monitor elders with chronic diseases inside and outside;
- Improve quality of life of the elder by giving more freedom and a feeling of safety;
- Prevent deterioration of patient condition with continuous support, guidance and health education.

The system is divided in three subsystems. The Home Subsystem that comprises sensors and other monitoring devices, to monitor the elder condition and behaviours inside home, and an interactive environment to provide health education. The Mobile Subsystem that consists in a group of wearable body sensors that collect important information about the elder vital signs, allied with a mobile phone that will allow to collect GPS data to ultimately analyse the mobility patterns of the elder. Lastly the Caretaker Site where all the information will be available to the care taker, as well as to health professionals, in order to allow the better care providing possible to the elder. [BLA+09] An overview of the system is presented in Figure 3.6.

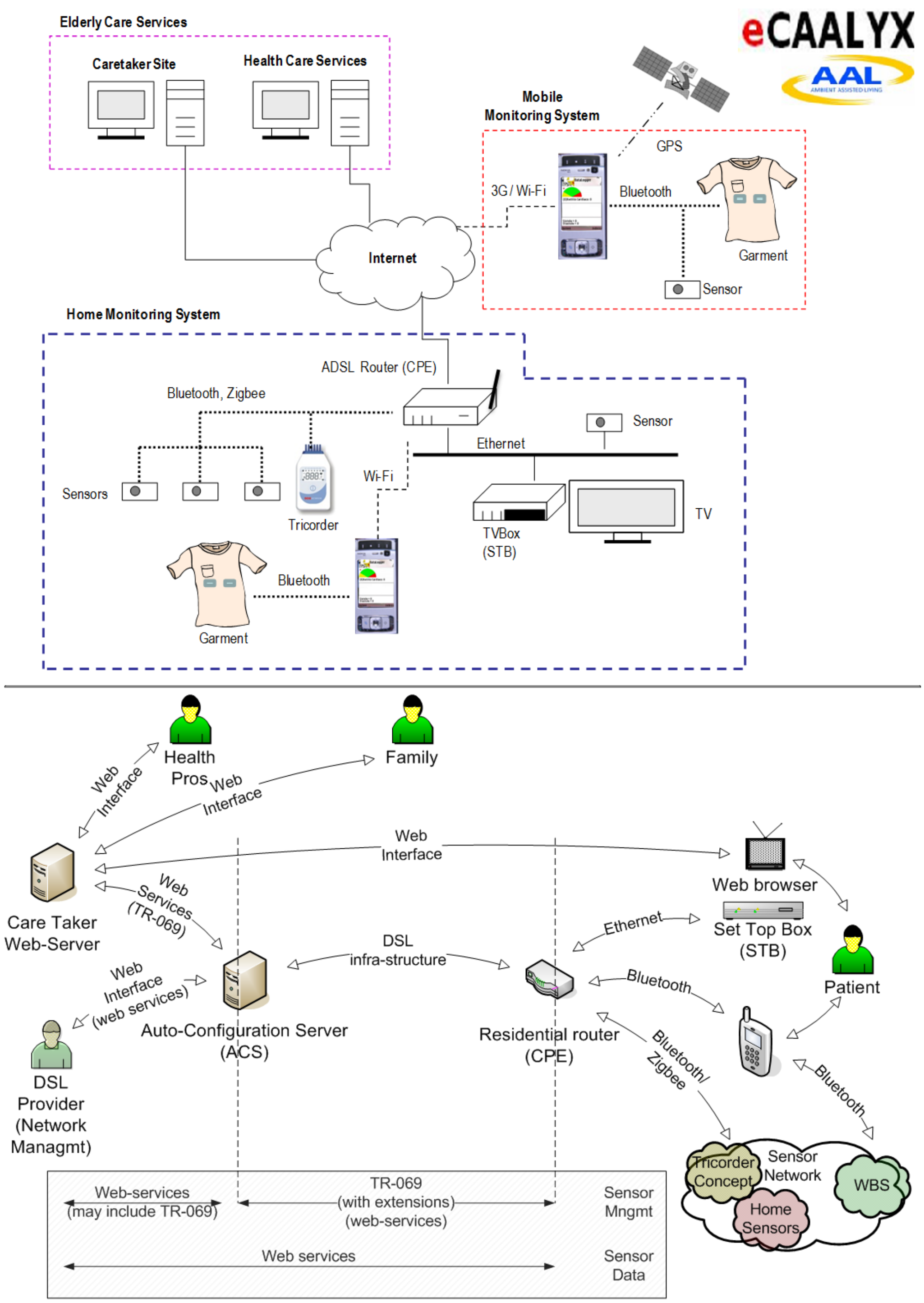


Figure 3.6: eCAALYX system diagram [BLA+09]

3.1.10 COGKNOW

COGKNOW is an AAL project mainly funded by the European Commission, that focus in helping people with mild dementia in carrying out their everyday living hopefully allowing them to live in their own home independently longer. [BLA+09] The main objectives of COGKNOW are in helping people with mild dementia in the following cognitive reinforcement fields identified by dementia sufferers and their caregivers, both in the literature reviewed and in workshops realized to that specific effect [MRBK+07], such as:

- Remember;
- Improve or maintain social contact;
- Allow leisure activities;
- Enhance feelings of safety.

In order to achieve these goals a user centred design/validation process was followed to develop a complete system. This system relies in both a sensorised home environment and a smartphone as a mobile sensing platform to collect information about the users and help them carry out their every day living. [DND+09] So the technological pieces that compose the COGKNOW system are:

- The Home Based Hub that gathers all the information collected by the sensors inside the house and sends them to the server and has a 17 inch touch screen that works as a graphical and personalized interface that conveys the information and messages to the user;
- The Mobile Cognitive Prosthetic works as a mobile version of the Home Based Hub, with which it communicates via Wi-Fi, fitted into a smartphone that may be used both inside and outside and also, when outside, takes advantage of the GPS functionality to guide the person home if he/she becomes lost;
- The Web-Based Server is the central repository of all information gathered, such as daily activities details and needs felt by the patients, and also allows caregivers to configure the reminders to be shown to the user;
- The Sensorised Home Environment that comprises all the sensors and actuators inside home that collect data about the user and sends it to the Home Based Hub.

All these parts come together as a complete auxiliary to the everyday living of a person with mild dementia, allowing for a better acknowledgement of their needs, difficulties and habits. [BLA+09]. A complete overview of the system parts presented above and how they come together can be seen in Figure 3.7

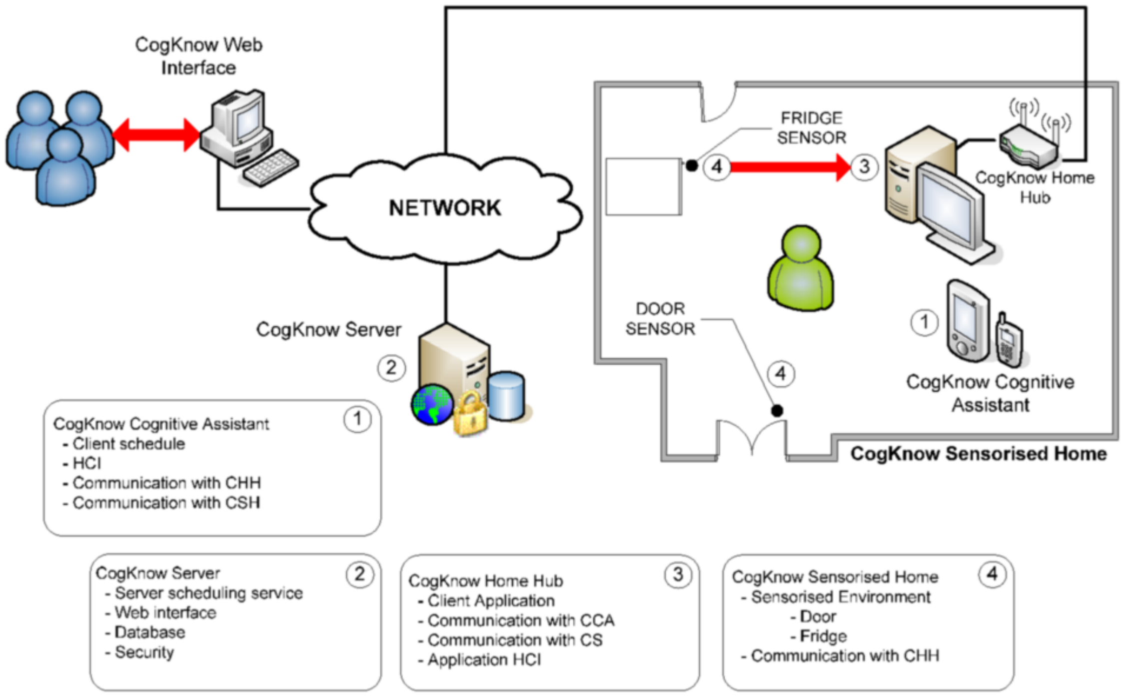


Figure 3.7: Overview of the COGKNOW system [MRBK+07]

3.2 Projects Overview and Comparison

In Section 3.1 were presented several projects that help and monitor people with different kinds of health problems, especially ones that imply cognitive impairments and difficulties like dementia. In spite of all projects having different objectives or different approaches to achieve similar ones, many of them have one or more points in common with each other. So it is important to understand exactly what each project covers and in what it fails to do when compared to one another. To understand the strengths and limitations of every project some comparative fields need to be used. One important field, as this project is centred in dementia is to see if the presented projects focus on dementia or are focused in monitoring a person for other reasons. From the literature was also possible to establish that outside wandering and getting lost was a problem common to dementia sufferers so tracking outside movement and detect possible outliers and interesting patterns are other fields. Finally, as some signs of dementia show themselves in everyday activities carried out at home it is important to collect data from various sensors and analyse and interpret it to find patterns and possible outliers. So taking this into consideration a table of comparison was compiled in order to allow a graphical and more straightforward analysis of the projects.

Looking at the Table 3.1 it is perceivable that none of the projects presented cover all of the fields considered. Even so there are clearly some more complete than others. The four projects that seem to be the more complete are eMotiva, The use of advanced tracking technologies for the analysis of mobility in Alzheimer's disease and related cognitive diseases, eCAALYX and COGKNOW.

	Sensor Data Collection	Outside Movement Tracking	Pattern Recognition	Outlier Detection (Warning Signs)	Dementia Centered
eMotiva	✓	✗	✓	✓	✓
Ginger.io	✓	✓	✗	✗	✗
Embedded Assessment of Wellness with Smart Home Sensors	✓	✗	✗	✗	✓
The use of advanced tracking technologies for the analysis of mobility in Alzheimer's disease and related cognitive diseases	✗	✓	✓	✓	✓
Electronic Tracking of Patients with Dementia and Wandering Using Mobile Phone Technology	✗	✓	✗	✗	✓
MONARCA	✓	✗	✗	✓	✗
SensCare	✓	✗	✓	✗	✗
AndWellness	✓	✗	✗	✗	✗
eCAALYX	✓	✓	✓	✗	✗
COGKNOW	✓	✓	✗	✗	✓
This Project	✓	✓	✓	✓	✓

Table 3.1: Table of comparison between referred projects

eMotiva [FLGGV+11] is a really complete project that collects data from several sensors allowing to establish behavioural patterns of the users and detect some changes to that behaviours, therefore outliers, that may be considered signs of dementia. In spite of this, because it is focused in institutionalized elders it does not cover the outside movement tracking. As aforementioned tracking outside movement is an important part of a tool that aims to monitor older persons that may present cognitive impairments and still live in their own homes in order to help achieve an accurate diagnosis of dementia.

The project presented in the article The use of advanced tracking technologies for the analysis of mobility in Alzheimer's disease and related cognitive diseases [SAF+08] is another very interesting and complete project. It does outside movement tracking and with that information movement patterns and outliers that may be related with dementia are uncovered. The downfall of this project is the fact that only focus on the outside movement. Doing so it does not collect data from other sensors that may provide valuable information about the habits and difficulties that the elder may face in his everyday living.

The European project eCAALYX [BLA+09] is an AAL solution design to monitor and improve the quality of life of elders. It collects data from various sensors and does also allow outside movement tracking. With the collect data behavioural patterns to the users are established. This project does not do outlier detection in the collected data and most importantly is not focused in dementia sufferers. Even so, it is a project that offers a guideline of how to develop this kind of solution but, as dementia has some specific problems and behaviours associated with it, it does not fit perfectly the problem at hand.

Finally other European project, the COGKNOW [MRBK$^+$07] is focused in people with mild dementia that live in their own homes. It tracks the outside movement of the elder as well as the movements and activities inside home. Furthermore, it tries to help the elder with some positive reinforcement in areas that are common problems to dementia sufferers. But as it aims in assisting the everyday living and not serve as a monitoring and diagnosis helper it does not do pattern recognition or outlier detection with the data collected.

It is easy to understand that there is a gap that needs to be closed, a system that monitors the elder person, gives assistance to health professionals in order to make an early and correct diagnosis and also allows caregivers to have more information that may help having some peace of mind. The project presented aims to close that gap, covering all these areas in order to be as useful as possible to caregivers, health professionals and obviously the elder.

Chapter 4

Outlier Detection and Algorithms

Outliers are exceptional or abnormal points in a data set that do not conform to the normal or expected behaviour or clearly fall in line to abnormal behaviour. As long as there has been data gathering outliers have existed in that data [SG11]. There are several reasons why outliers may appear such as human error, faulty machinery, environment change or malicious intended activities.

In the beginning outlier detection was used to remove these points, considered as noise, from the dataset in order to avoid their influence in the results. Nowadays noise removal and outlier detection can be seen as two different disciplines. Even though data noise and outliers are usually very similar they can be approached in different ways. Noise points do not offer interesting or important insights into information and they usually get in the way of a correct and precise analysis of the data and can corrupt the results. Outliers on the other hand can present interesting and unique insight about what is happening, if something has changed or if a problem is occurring. These exceptional points may contain valuable information capable of indicating malicious intent, such as hackers trying to access and download information from a server or a stolen credit card being used, human errors, that may jeopardize a project, and even machinery malfunction that if gone undetected could influence the quality of a product that comes out of a production line. Outlier detection can be used in a myriad of other domains, in particular to detect abnormal behaviour of a person that may indicate the presence of an illness [CBK07].

Outlier detection techniques may be divided into several groups depending on the criteria. We can divide them by the type of supervision they require, arising the supervised, semi-supervised and unsupervised outlier detection algorithms. Supervised techniques assume that training data for both normal and outlier points is available and usually predictive models are used to categorize new data as normal or outlier. This is possible because this techniques have a clear notion of what is the normal data and the outliers. The biggest problem with supervised algorithms is that training data can be expensive or even impossible to get, reducing the viability of this approach in some cases [AZL06].

Semi-supervised techniques are similar to supervised ones but only need labelled data for the normal instances, everything that falls outside of this normal data is considered as belonging to some

other class, usually outliers. These techniques arise from the difficulty of, in some cases, having labelled data for the outlier class. But even avoiding the need of labelled data for the outlier cases, the need of labelled data presents the same problem as in the supervised techniques as obtaining this data is usually complicated.
Finally there are the unsupervised outlier detection techniques that do not require any type of labelled training data and therefore are more widely applicable. In this case it is assumed that the normal types of behaviour are the ones that are more frequent and any rare occurrence is considered an outlier. The biggest problem with this type of techniques is that they present a much higher false outlier rate because some of the assumptions made may not be true [CBK07].
These techniques or algorithms can also be divided by the discipline that gives basis to its concepts. These disciplines range from very broad ones like Statistics or Information Theory to sub-fields like clustering, nearest neighbour or classification. And for each of these types of approach exists a number of subdivisions and algorithms implemented. The type of approach chosen is dependent on the problem, how it should be addressed and what information that available for analysis.
Besides all the work that is behind finding and understanding outliers and how to detect them it is important to take into consideration the possibility of *Concept Drift*, especially when dealing with applications that want to understand and create a representation of the real world. Sometimes there is a fine line that separates outliers from a drift in the concept that is being analysed. For example, if a person changes its living location, the first data analysed may be perceived as an outlier, as the person may not have been in that place before. Any application that wants to deal with real world data must be able to adapt quickly and understand when something is truly an outlier or if a Concept Drift has occurred. [Tsy04]

4.1 Methods and Algorithms

As aforementioned there are several types of outlier detection algorithms and methodologies that fall in many different categories. This section will introduce some of them, their strengths and weaknesses and in what cases they might present themselves useful. Some of the algorithms and methodologies presented will be used in the development of the work described in this dissertation and some others may be used in some possible future work.
From the methods and algorithms described below the ones used in the implementation of this project were:

- Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm was used to deal with spatial information and discover outliers in the movement outside home.
- Dynamic Time Warping algorithm was used to compare time series and perceive if they are similar or if on the other other hand they have some meaningful differences.
- Sequential Pattern Mining, in specific the Generalized Sequential Patterns algorithm, was used to discover the sequential steps taken by a user when dealing with his everyday life activities

- Statistical Methods were used to establish the normal behavioural patterns of the user and also to discover some outliers that clearly deviate from the normal pattern.

4.1.1 Statistical Methods

Statistical approaches to outlier detection were the first ones to be used in outlier detection. [HA04] These methods firstly arose in the 19th century from the necessity felt by statisticians of finding and removing abnormal points that introduced bias into the statistical analysis of their observations. [CBK07]. A statistic approach to outlier detection basis itself int the notion that the data is produced by a stochastic model and any data that is unlikely to be generated by that method may be perceived as an outlier. A very interesting example of this is the one used by Barnett that explores a the judicial case of Hadlum vs Hadlum (1949) where the child was born 349 days, approximately 50 weeks, after Mr. Hadlum left for military service. Barnett conducted observations of 13634 gestations period and found that the gestation period, that in average is of 40 weeks, followed a Gaussian distribution, as is shown in figure 4.1, and that the birth of Ms. Haldum was clearly an outlier in this case. [Bar78]

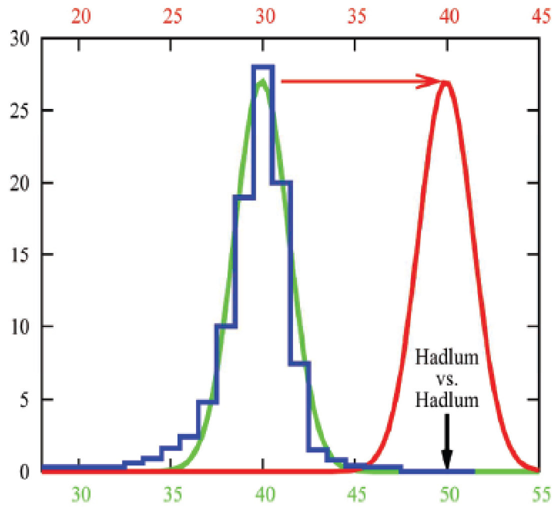


Figure 4.1: Representation of Barnett findings and Ms Hadlum pregnancy

The basic statistical models to outlier detection may only rely in simple calculations like averages, means, standard deviation and box plotting. [HA04] So it is easy to understand that statistical approaches offer a simple and easy to understand way of finding outliers, as it follows closely the idea that these are points that do not relate with others. The biggest problem with these methods is that values like the mean and standard deviation are very sensitive to outliers and these values need to be computed from the whole dataset, that may contain these outliers.

4.1.2 Density-Based Spatial Clustering of Applications with Noise (DBSCAN)

DBSCAN is a density-based clustering algorithm that is able to discover clusters of arbitrary shape and noise from a set of information and that was first presented in 1996 [EKSX96]. Since then

many variations of DBSCAN have been introduced in order to adapt it to concrete problems or set-ups and also trying to diminish its complexity [AAS10]. Even though the primary objective of this algorithm being the clustering of data, its capability of detecting noise, i.e. outliers, makes it a good algorithm to highlight those points.
DBSCAN relies only in two parameters to construct its clusters and find the noise points, Eps and MinPts. Eps is the radius where the neighbours of a point are searched. MinPts is the minimal number of neighbours that a certain point must contain in order to be considered a core point of cluster. The algorithm starts arbitrarily in one of the points and checks its Eps neighbourhood, this is how many points are inside the Eps radius of that point. If the point has enough points in its Eps neighbourhood, more than the value of MinPts, the algorithm will try to expand that cluster, otherwise that point is marked as noise. This does not necessarily means that in the end this point will be considered as noise because it may be density reachable from another cluster. Being density reachable means that a point may not be a core point of the cluster but is in the Eps neighbourhood of a core point and therefore being a border point of the cluster. The border points, the ones that are not core points of the cluster, are found in the expand cluster phase. In this phase, after establishing that a point is a core point, every neighbour of that point is visited, if it contains enough points in its Eps neighbourhood is also labelled as a core point, otherwise it will be labelled as a border point. In the end we have all points labelled as belonging to a cluster or noise points [EKSX96]. In figure 4.2 it is possible to see two clusters and some noise points resultant of a DBSCAN algorithm execution. Due to their shape this type of cluster would not be well handled by some other clustering algorithms.

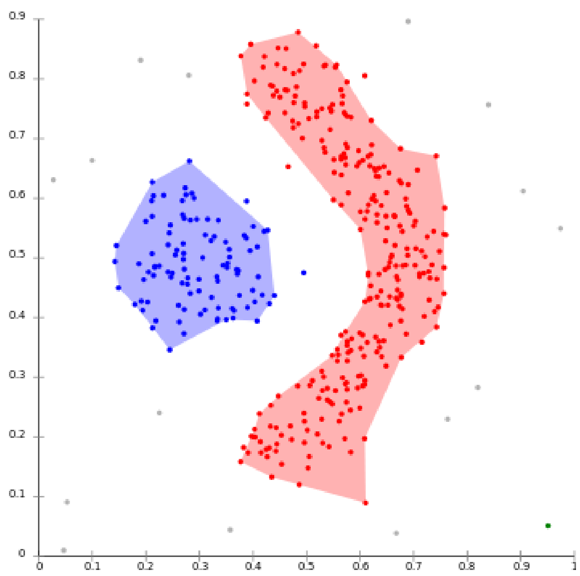


Figure 4.2: Example of two clusters created by a DBSCAN execution

So the biggest strength of DBSCAN when compared with other clustering algorithms is its ability of detecting arbitrarily shaped clusters and also discover how many clusters exist without any a priori information about this. The main problem when using DBSCAN is its complexity that may reach $O(n^2)$ [AAS10]. It is possible to see that DBSCAN can serve as an outlier detection algorithm if instead of eliminating the noise points, we save and analyse them.

4.1.3 K-means

K-means is, like DBSCAN, a clustering algorithm that has several variants but its basic implementation is also known as Lloyd's algorithm. [KMN+02] The name K-means is due to the way this algorithm finds clusters from a variable k that represents the number of clusters to be found. [Mac03, p.284] In the first step of the algorithm are generated k means, that ideally will be the center points of the clusters. This first means can be generated in several ways, but the basic approach is to generate them randomly. Having the two means the algorithm then becomes a two-step iterative process. In the first step the points are associated with one of the k means, the one that is closest. Then the means are updated to be the centroid of the clusters created by the previous step. The algorithm is finished when there are no changes to the means positions and their neighbours. [KMN+02, Mac03] In figure 4.3 it is possible to see a graphical representation of K-means steps.

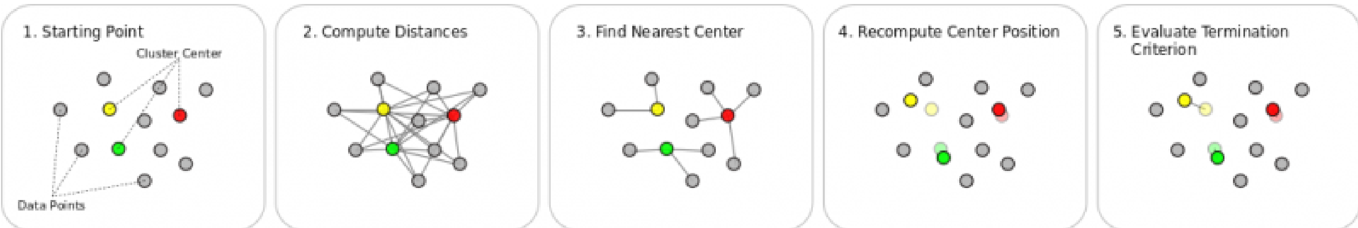


Figure 4.3: Example of the K-means execution steps

From the description presented above is simple to understand that one of the biggest limitations of K-means, besides its complexity, is that it needs to previously know how many clusters we pretend to find. This limits the use of this algorithm, in its basic form, to only some specific use cases where the number of clusters is known beforehand. In spite of this there are some methods of finding the value of k computationally using for example the Average Silhouette Width Criterion that is computed after the clusters are found. If the value of this criterion is optimal then the value of k is a good one, otherwise other must be chosen and the process repeated. [CH06] This further adds to the complexity of the execution which is undesirable. Also, as it was presented, K-means does not offer the possibility of finding outliers in the data. But there are some ways in what K-means can serve this purpose. One approach to achieve this is presented by Hautamaki [HCK+05] where after the clustering of the points, its is calculated an outlierness factor to each point. This factor is calculated by diving the distance of a point to its mean and then dividing it to the maximum distance of a point to its mean. The points with an higher outlier factor may be considered outliers.

4.1.4 Outlier Scoring based on the K nearest neighbours (KNN)

The K nearest neighbours have a very self explanatory name, as it relies in exactly that, the *k* neighbours that are least distant from a point. KNN as several applications in data mining, one of them being classification, where having labelled data an entry is classified as being equal to the majority of its neighbour points. [WKRQ+07] In spite of this, the basic idea behind KNN can be easily used as an outlier detection auxiliary. One interesting way of doing this is presented by Ramaswamy [RRS00] where the K nearest neighbours are used to calculate an outlier factor for each point. To calculate the outlier factor is used the distance to kth neighbour. A similar approach was followed by Angiulli [AP02] but int this case the outlier factor is calculated using the sum of the distances of all *k* neighbours of a point instead of only the furthest one.
The limitation of this approach is easy to perceive. The number of *k* neighbours is an essential variable and therefore having to defining it beforehand may cause unwanted results. If a too small value is used small clusters of outliers may go unnoticed. On the other hand if this value is too big some points may have undesired points int its *k* neighbourhood and therefore being considered as outliers when that may not be the case. [WKRQ+07]

4.1.5 Local Outlier Factor (LOF)

The notion of Local Outlier factor was first presented by Breunig et al. [BKNS99, BKNS00] as an alternative solution to the global and binary outlier detection methods, the only existent until there. With their work the notion of scoring a data point instead of just being an outlier or not was introduced. This method is, in its essence, similar to the Outlier Scoring based on the K nearest neighbours, it in fact uses KNN in one of its steps. The biggest difference in LOF is the notion of locality. The authors justify the importance of this with a simple example of a data set that contains two clusters of different densities, like the one in figure 4.4. Using a global notion of outlier is easy to understand that the points in cluster C1 would have a similar outlier score as point o2 as both K neighbourhoods are at a similar distance. But this clearly does not represent the reality as point o2 may be an outlier point close to a very dense cluster an points in C1 are points inside a less dense cluster.

To achieve the outlierness score based on local density the LOF algorithm uses three measurements. First the K neighbourhood distance (k-distance) that is simply calculated with K nearest neighbours algorithm already presented. Then is also calculated the reachability distance of an object, that works as a smoothing factor. If the two points are distant the reachability distance is simply the distance between them but if they are close to one another the reachability distance will be simply the k-distance of the object. Finally is calculated the local reachability density, that is simply the inverse of the reachability distance of a point. Having this three variables is then computed a score of outlierness for each point.
The biggest strength of this approach is its capability of accurately detect outliers in data sets that may have distinct clusters with different densities, and give a score to each point. It is the possible to consider outliers the *n* points with higher outlier score or the points that have an outlier score

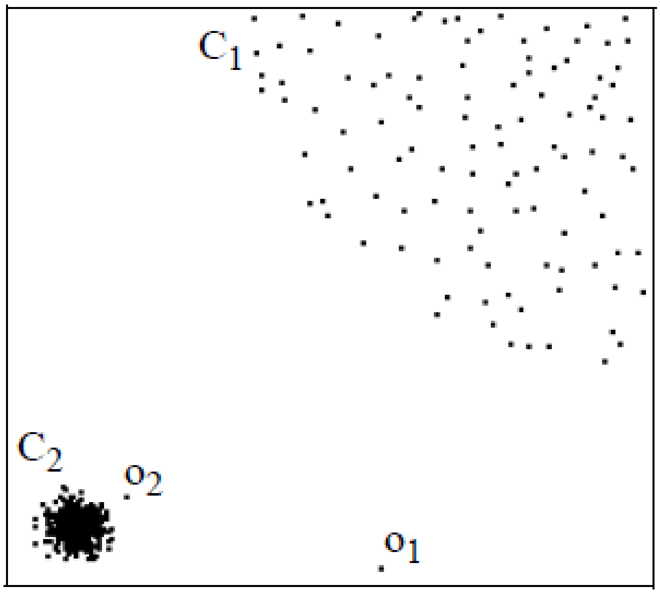


Figure 4.4: Example used by Breunig et al. [BKNS99, BKNS00] to explain the importance of the notion of locality

bigger than a given threshold. As the LOF algorithm is designed to compute an outlier score to every point in the dataset its time complexity is O(n^2) which may become a problem when working with big dataset. [AJK10]

4.1.6 Angle-Based Outlier Degree (ABOD)

The Angle-Based Outlier Degree outlier detection algorithm, presented by Kriegel et al. [KShZ08], detects outliers using primarily the direction of the distance vectors between the points instead of using the distance itself. The rationale behind this approach is that in high dimensional data spaces the distance between the points becomes an irrelevant value and therefore outlier observations may easily go undetected. On the other hand the direction of the vectors will enable to discover outliers even in high dimensional data spaces. This is because the directions of the distance vectors of points in the center of the clusters will always have very different angles, even border points will have some variation in their values, but outliers will have little variation in the angles, as it is easily perceivable in figure 4.5. The intuition that lead to this algorithm becomes clear when we look into figure 4.6.

For each point in the data is computed an Angle Based Outlier Factor (ABOF) by calculating the difference of the scalar product of the distance vectors to every other pair of points in the dataset normalized by the quadratic product of the length of the difference of the vector. Here is possible to see that distance plays a part in weighting the calculation, a bigger distance makes the angle weight less, because the angle variance is stronger for a bigger distance, and this could lead to having some outliers going undetected.

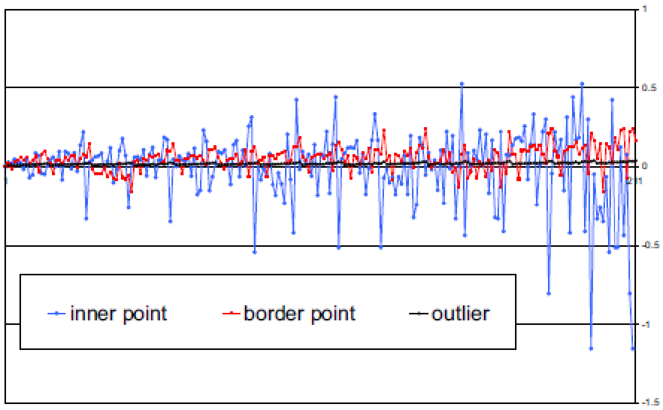


Figure 4.5: Variation of the angle values [KShZ08]

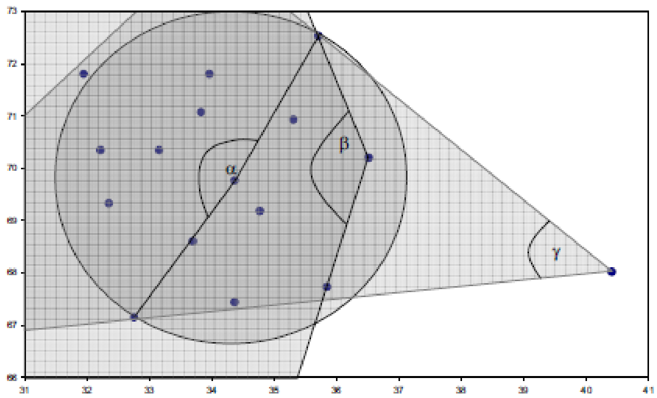


Figure 4.6: Rationale behind the ABOD algorithm [KShZ08]

Overall this is a distinct approach to outlier detection, using angles instead of distances, and correctly detecting some outliers that some other algorithms may struggle to find. One drawback is its complexity, that is O(n^3), because for each point every other pair of points needs to be visited. To overcome this some variations of the basic implementation of ABOD are suggested by the author.

4.1.7 Dynamic Time Warping (DTW)

The Dynamic Time Warping algorithm was first introduced in 1983 [KL83] to allow comparison between similar time series where one of them may have a slight shift in the time series in a simple and intuitive way. This shifting capability allows this algorithm to compare time series of different lengths or with missing records. Time series analysis, in particular using Dynamic Time Warping, has been used in a number of fields such as speech and gesture recognition, robotics and time series streams. [SC07, LWLW08]

A basic way to compare two time series is simply to use the Euclidean Distance from the nth element of the first series to the same nth element of the second. This introduces some limitations as the time series need to have exactly the same length and two similar time series may appear different if one is just slightly shifted. To overcome this instead of comparing only the elements of the same index, the Dynamic Time Warping Algorithm, for every element in a time series X, finds the element of minimal Euclidean distance in the whole time series Y, like is possible to see in figure 4.7 where the two time series are somehow similar but a shift is needed to further comprehend that similarity. The results returned from the execution of the algorithm are a warp distance, as in the distance between the two time series, and the warp path, this is which index from time series Y was matched with which index of time series X.

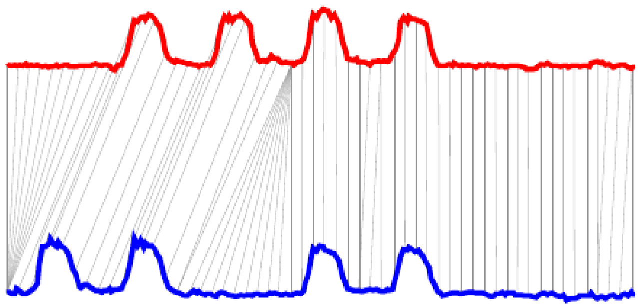


Figure 4.7: Warp path between two time series

As many as the algorithms presented, the major weakness of the Dynamic Time Warping algorithm is its complexity that is $O(n^2)$. To overcome this some work has been done, like the one presented in [SC07] where, instead of comparing to all the indexes of the time series, a window is used to reduce the number of records to be compared and some indexing is also applied.

4.1.8 Sequential Pattern Mining

Sequential Pattern Mining consists in analysing and discover frequent sequences of events from a set of information. It has been used in a number of fields such as DNA sequences, Web click-streams or customer shopping sequences. There are several specific algorithms to discover sequential patterns like Generalized Sequential Pattern (GSP) [SA96] designed to be used in static databases, SS-MB and SS-BE [MDH08] thought to be used on streams of data. [HCXY07]
Discovering these sequential patterns may offer insight to what are the normal sequences in a specific case and easily detect if new, and possibly abnormal, sequences start to occur. Is in this way that Sequential Pattern Mining may be useful as an outlier detection method. And if a new pattern starts to become one of the most frequent then this new pattern may not in truth be an outlier but a Concept Drift instead. [Tsy04]
It is easy to understand the multiple uses of mining this Sequential Patterns as they can offer new different insights that would go unnoticed if the data entries were analysed one by one instead of looking at them as a sequence. Furthermore, the possibility of adapting this Sequential Pattern Mining to streams of data opens a whole new range of uses to this data mining technique.

Chapter 5

Project Specification and Implementation

As referred in Chapter 1, our world is ageing, and at a fast pace [CDRV09], which has a direct influence in the predicted evolution in the numbers of dementia sufferers. [OI12, p.4,p.12] Therefore it is important to focus in understanding and developing solutions to help these people.
There has been work made in several fronts of this area, like the ones presented before. But, as aforementioned, it still exists a gap that needs to be filled. It was not possible to find a complete system focused in monitoring and help people that may suffer from dementia by helping achieving an accurate diagnose through the detection of possible warning signs in the behaviours of those people.
It is also of vital importance helping persons already diagnosed with some sort of dementia by monitoring the development of the disease. The works presented in Section 3.1 focus in different areas. One is Ambient Assisted Living that is focused in improving the quality of life by simplifying some day to day essential tasks and reinforcing positive behaviours in order to fight the progress of dementia. Others focus in giving peace of mind to caregivers by allowing them to monitor the elder when he is alone. And some give their contribution by trying to help to offer the possibility of getting an as early as possible diagnose. This is done by collecting data that represents the behaviours of a person and by interpreting those behaviours. This interpretation may be valuable to catch some signs that may be indicative of dementia. Furthermore these solutions offer to health professionals the possibility of accessing large amounts of data in an organized fashion to simplify its analysis.
This project proposes to focus in supporting dementia diagnosis and monitoring. But, first of all, it is important to understand that the diagnosis of dementia is a complex and not yet linear process. There are not specific guidelines to what signs, and in what degree, may be considered dementia. Because of this any system will always have work as a support to gather information to help health professionals. Obviously the system can do more than this, analysing the data, understanding what is important in order to only show information that may relevant and even try to interpret the data in order to find some signs that may indicate the presence of dementia and thus giving a warning to

the person that a visit to a professional may be important in order to understand what those signs might mean. And it is exactly this that is proposed to be achieved in this project. We propose to design a system that gathers data from various different sources such as GPS tracking data, home sensors and even data from calls, messages or emails, and with all the information gathered and organized, interpret it and warn if possible signs of dementia are identified. Furthermore, all the data collected may be of high value in the process of diagnosis and monitoring carried by professionals specialised in the area.

The main difference between this project and the available solutions, see Section **??**, is the ability of collect and integrate data from different sources about a person living alone. This is important because elders tend to resist being left in nursing homes. Such a system may offer an opportunity to extend the time that the person stays at home by supporting an early diagnose and monitoring behaviours. This will help health professionals in their diagnosis and monitoring of the elder. Furthermore will also give some peace of mind to family members and caregivers by giving them the opportunity to have a detailed view of the behaviour of the elder and be warned in case some risk behaviour appears.

And because nowadays there are a large number of older people living alone at their homes, especially in big cities, and their family members do not have the time to be with them at all times, due to overloaded schedules. This gives them the opportunity to remotely follow the behaviours of the elder.

5.1 Project Specification

This section is aimed at giving a clear view of the system and its specification. First of all it is important to define the scope of the project. During the development the focus was given to the specification and implementation of the back-end processing and storage of the data and its presentation in web visualizations as well. This means that it was not an objective of this project to develop the sensing platforms such as smart home environments or data collection through the usage of smartphones. The goal of the project is to design a generic system, capable of receiving different types of streams of data, analysing and storing it, and then present the results to the users in a simple and easily understandable visualization. Another aspect taken into consideration was that the system should be modular in order to be simple to extend, making easy to add new functionalities and analysis of new types of data.

The decision of focusing the project on the analysis, storage and presentation of the data was made because the sensing structures and data gathering platforms already exist, like the ones described in Section 3.1. What was important was to develop a system that was capable of receiving, interpreting and give meaning to that data in order to add value to it as an auxiliary tool in dementia diagnosis and monitoring. Furthermore this tool also aims to promote an early help seeking in those that may be presenting some early signs of dementia, that easily go unnoticed by the elder and his family.

Such a system, when allied with all the data gathering platforms, will be continuously receiving streams of information for each user. It becomes clear that the system will have to deal with great amounts of data. Only some hundreds or maybe a few thousands of users, continuously generating data creates a big stream of information that has to be analysed and made ready to be presented to the user as soon as possible, in order to give the possibility of looking to the most recent data collected as soon as possible. Making the system capable of handling this flow of data was other point taken into consideration.

5.1.1 Features specification

In order to have a specification of the project it was necessary to specify all the features the system should include in order to fulfil all the objectives. Furthermore some non-functional requirements were taken into consideration to make the best of the features proposed.
Starting with the specification of the features and taking into consideration the system envisioned in this project it must be able to:

- Handle different types of streams of data such as GPS reads, home and smartphone sensors data and others that may be considered relevant (i.e. call history, browser history and messages);
- Be easily extensible, allowing new types of data and analysis to be added;
- Provide real-time analysis of big streams of data;
- Uncover behavioural patterns in the data analysed;
- Discover possible outliers in the data received;
- Assess if the outliers detected may be related with possible signs of dementia;
- Store all the pertinent data received and generated by the analysis process;
- Present the analysed information to the user in a simple and clear way;
- Allow users to check data from different periods of time, with different levels of granularity (in the cases that this may apply).

In order to make the best of all the functionalities made available this system the following non-functional requirements were taken into consideration:

- Usability
  - The design of the system, specifically the web user visualizations should follow existing usability guidelines. As the focus of this system is to help older people, specially persons who may suffer from dementia or present some cognitive impairment, the visualizations need to be clean, without any unnecessary options in order to not become a source of frustration for its users. Ideally different visualizations should be designed

for doctors, caregivers, family members and for the elders as they will may have different objectives in the use o the system.

- Reliability
  - Since the system aims to be an aid for the diagnosis of dementia as well as work as a monitoring tool to help track the development of the disease and give some peace of mind to caregivers and family members it should be as reliable as possible. It must not present false information to its users and specially it should never overlook a possible sign of dementia. It will be preferable to highlight a false positive than let a possible warning sign go unnoticed.
- Robustness
  - In order to maintain the information always updated the system should be robust. In case of failure in the visualizations the user should be clearly informed that an error occurred instead of failing silently. As a complex system, when run into production, the whole system will need to be managed and monitored. If an error in the backend occurs the system manager should be informed with an explanatory message that allows him to quickly solve the problem.
- Error Recovery
  - The system should, whenever possible, recover from any error that may occur. As this is a complex system with several pieces working together and relying on each other any possible error that happens the system must be able to recover from it as fast as possible.

5.1.2 Usage Scenarios

To achieve a well thought, efficient and robust design for the system it was important to first define how it will work, what basic usage scenarios will be implemented in order to create a design that is well adjusted to them. The following list of usage scenarios was thought as a basis implementation of functionalities of the system:

- **Scenario 1:** An elder has a set of places that he normally visits. Suddenly he may start to go to new places that are not normal. This may be a sign of disorientation, getting lost or visiting places that are not part of his day to day life, which are common signs of dementia. Is important to take into consideration that this may also be a concept drift, as in the case of a family member of the elder that changes the place he lives and he starts to visit his family member there. If this is the case this place through time will be visited often and this change should be disregarded, as an obvious concept drift as occurred.
- **Scenario 2:** An elder may get disoriented in his normal day to day courses. For example, he may be going to the grocery store and all of a sudden becomes disoriented and not knowing

exactly where he is. He may start wandering around, looking for something to help him understand where he is. As a consequence of this the duration of this regular trip will differ from the normal time and new travelling paths may arise. The system should be able to detect this and highlight it as a disorientation period. Here once again a concept drift may occur, a street that is closed may force the elder to take different and more timely paths so if this becomes usual, these changes should also be disregard as outliers and attributed to the concept drift.

- **Scenario 3:** A person usually leaves the lights, taps and appliances turned off when he leaves home. Forgetting a light turned on every now and then is not cause for concern. But if a person starts to leave home and leaving all their home lights turned on this may be considered a sign of forgetfulness and therefore a sign that something more could be happening.

- **Scenario 4:** An elder usually has some medication that he needs to take. For example, if every day he takes some medication when he wakes up plus some more before going to bed and, at some point, starts to regularly forget taking the night time medication this can be considered a sign of worrying forgetfulness.

- **Scenario 5:** A person has a normal pattern of inside movement. Usually there is more movement in the morning, at lunch and dinner times and before going to bed. Normally night time is more quiet with little movement or none at all. If this pattern suddenly changes and for example starts to appear a lot of movement during the night time this may be a sign of night wandering, what is common in people with dementia and is usually cause for concern because the person can be disoriented and easily get hurt or even leave home and seriously getting lost without noticing it.

- **Scenario 6:** A person has a number of times that he usually leaves home. For example in a week a person may go outside 10 times. If this number starts to steeply decrease and a person starts to leave the house only 1 or 2 times per week, or even none, this may be a sign of social withdrawal and therefore cause for concern.

- **Scenario 7:** The mean time spent in some house divisions may be related to the time spent carrying out some tasks. For example the time spent in the kitchen before lunch and/or dinner may be seen as cooking, and time spent in the bathroom in the morning or night may be perceived as taking a bath. As cognitive functions decline more difficulty is expected to be associated to these tasks, and therefore more time spent in carrying them out. So if a person usually took an average of 30 min to cook lunch and that average time increases, it may be due to increased difficulty in carrying out the task. Obviously this increase in the time it takes to complete a certain task can be due to other difficulties and not cognitive impairment, but it is up to the doctor, talking with the patient to understand what is the underline cause.

These were the usage scenarios taken into consideration designing the system but, as mentioned before, the system must be able to adapt to new sources of information and other scenarios, and therefore these were just the guidelines to design the basic architecture behind the working system.

5.1.3 System Architecture

After defining what are the requirements that the system needs to fulfil it was vital to carefully design a system which architecture would allow those requirements to be achieved. As mentioned before the effort was to develop a system that receives, analyses and stores the data to then present it in a meaningful and simple way to its users. In order to do this the architecture presented in Figure 5.1 was designed where it is possible to see all of the different pieces that compose the system and how they work together.

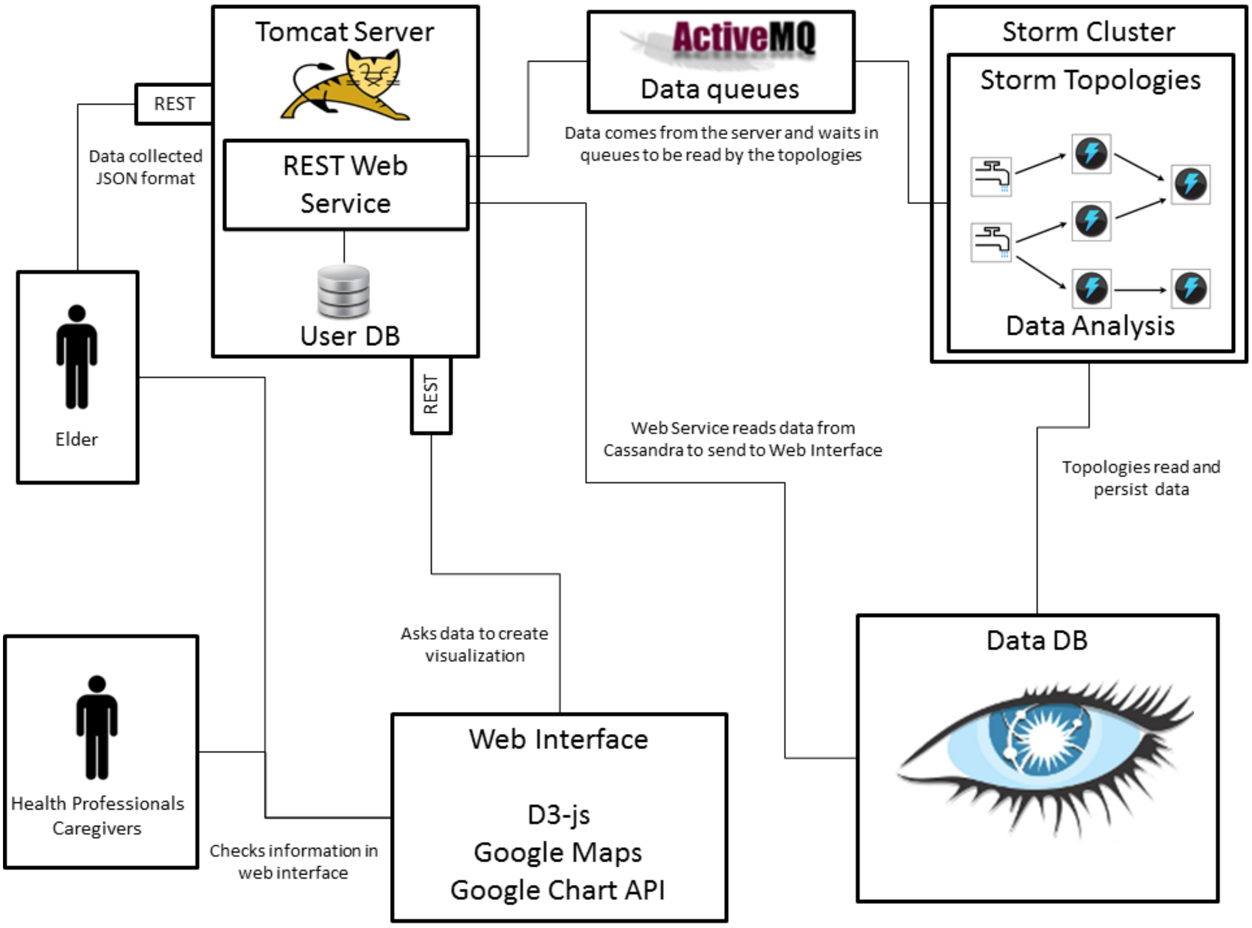


Figure 5.1: Overview of the proposed system architecture

Looking at the top left end of Figure 5.1 we see the server, which contains the REST web service. REST stands for Representational State Transfer and is a type of architecture for distributed systems. It is used to implement communication between clients and the server. This allows the clients to make requests to the server that processes that request and gives an appropriate response. In this case the REST web service serves as a connection point for all the system. It is the first entry point of all the information. This is done by the several sensing platforms sending requests to the service with the data collected. This request is then processed by the server that takes the information and forwards it to the corresponding queues. Different types of data, that will go through different types of processing, are put in different queues in order to allow simple access to the correct information at all times. Furthermore the web service works as linking element between the web visualization and the database where the behavioural data collected and information resultant from its analysis is stored. The web page that will show all the information sends a request to the web service asking for the data. This request is processed and the database is queried for the pretended information. This informations is then sent back to the web page to be presented to the user. The server will also contain the user database, where basic information about the users is stored. Additional information about the server, and the REST web service in particular, is provided in Section 5.2.1.

Now looking at the Storm [Mara] cluster, visible in the top right end of Figure 5.1. Storm is a distributed real time computation system that facilitates the processing of streams of data. The Storm cluster is where all the data processing, analysis and storage decisions are made. All this process is done in a distributed fashion in order to allow quick processing of large streams of data. The data that was forwarded to the queues by the REST web service will be read by the components of the cluster. This components can be Spouts or Bolts and together they form topologies that are the graph of the processing of information. Spouts are responsible for fetching the information from the queues. The Spouts will easily have access to the correct information as this is previously processed and forwarded to the correct queue by the REST web service. On the other hand Bolts are responsible for doing all the data processing. After all the analysing steps are completed the relevant information is stored. Further explanation about the Storm cluster its internal components and how the data analysis is made will be provided in Section 5.2.2.

The data database, visible on the bottom left of Figure 5.1, is where all the data collected and further information resultant of its analysis is stored. This database is built with Cassandra [Fou], a NoSQL and distributed database. Such a database was used in order to allow distributed and replicated storage, as large amounts of data are expected because for each user continuous streams of information will be produced at all times. All that information needs to be stored and available as it may be relevant, not only to be analysed as is, but also to be a point of comparison for future information that may appear. This part of the system is detailed in Section 5.2.3.

Finally the web user visualization. This is where all the relevant data can be checked by the users. The visualizations will consist in maps with the collected information, charts and other graphical representations of the information. This is the only part of the system visible to the user and should be kept as simple and clean as possible to allow an easy access to all the vital information. This means not having more information than what is strictly relevant, in order to not loose focus. The implementation details of the visualizations are given in Section 5.2.4

5.2 Implementation

After a thorough specification of what the system should do, how it should do it and defined a good design to achieve those goals, some functionalities of the system were implemented. It was necessary to test the system with different types of information that could give insight into possible dementia signs. Due to time constraints it would be impossible to implement all the functionalities envisioned. Therefore we decided to choose a set of usage scenarios that would use different types of data and give insight to different types of dementia signs referred in bibliography and presented in Chapter 2. For these reasons it was decided to have a case where outside movement was tracked, other that analysed movement inside home, one where forgetfulness signs would be detected, a scenario where social withdrawal signs would be visible and finally one that could give insight to how a person tackles his every day tasks. This way it was possible to understand if all the design decisions made would work and if this system could be in fact a valid monitoring tool and an aid in the diagnosis of people with dementia. In the following sections the major components of the system developed will be presented in detail, justifying the decisions made and giving an understanding of how the system works. All the work implemented and described below was developed in Java.

5.2.1 Server and REST Web Service

The Server is composed by the REST web service and the user database. As it was stated in Section 5.1.3, the REST web service will serve as a vital linking point for all the system being a bridge through which all the information flows.
All the information enters the system through the REST web service ideally in JavaScript Object Notation (JSON) format. JSON is a lightweight data interchange format, readable by humans and easily parsed and produced by machines. Although it would be possible to convert different types of data to JSON in the web service by adding some processing to the data it is preferable to receive all the data in the same format. Different types of information have different paths to which they should be sent, so a GPS read will be sent to a different URL than a movement read. When a piece of information is received the web service forwards it to the corresponding queue in order to make it available to the correct Storm topology. Before forwarding the information to the Storm topology the web service needs to validate it by making sure that it belongs to a registered user in order to avoid unnecessary processing and storage of pointless or erroneous data.
The web service also works as a link between the web graphical visualization and the Cassandra database. All the information required to design the charts, maps and other graphical representations is requested to the REST web service. Using the data that comes with the request the web service then queries the Cassandra data base to collect the pretended information. That information is processed and the adequate JSON response for each visualization is created, avoiding unnecessary work on client side Javascript. Examples of the JSON produced by the web service and sent to the different visualizations can be seen in Appendix A In order to create accurate responses in the JSON format it was used the Google Gson library [SL] to serialize and de-serialize

all the information. This way errors, common when creating the JSON string manually, were easily avoided and all the process was simplified.

5.2.2 Storm Topologies

The Storm topologies, contained in the Storm cluster as presented in the System architecture overview given in Section 5.1.3, are where all the data analysis and storage orders are produced. The topologies are the graph of execution for the information that comes in where the nodes contain all the process logic and the edges show how the information should flow between nodes. [Marb] The nodes in a topology are either Spouts or Bolts. Spouts are where the information comes in and then addressed to the respective Bolts. In this case all the spouts read the messages stored in the queues that were forwarded by the Server, but other sources of information can be used given the correct configuration to the spout. The Bolts are where all the data processing and analysis is made and possibly forwarded to another bolts, if the more execution steps are required. Several instances of a Spout or Bolt can be running at the same time what allows to speed up the execution process what is important in order to have real-time processing without significant delays. The edges of the graph are also referred to as stream of data inside the topology. The way the stream flows between nodes and to which task of a Bolt the information should go is fully configurable using some of the groupings provided by Storm. For this project the most useful one was the grouping by a field allowing to assign each user to a specific task of a Bolt guaranteeing that the data of one user is always treated by the same task, avoiding undesirable race conditions and making possible to treat data from different users in a parallel way.
Using Storm it becomes really simple to extend the whole system because if a new type of data needs to be analysed the only thing that needs to be done is to create a new Topology and then submit it to the cluster, with no interference with the topologies already running. Another positive aspect about Storm is the fact of being simple to configure and monitor, as before storm developing a real-time system would involve a complex network of workers and queues. Furthermore systems based on this previous design were very difficult to scale and not fault tolerant. Storm overcomes all these limitations, being simple to extend and fault tolerant as proven by many production clusters of storm used by several companies. [Mara]
In the following sections will be presented the Topologies developed in order to further understand the data processing behind the system.

Outside Movement Topology

This topology was developed to analyse the data concerning the outside movement in order to find possible location that are not normal places for the elder to visit but do occur. In Figure 5.2 is possible to see a graphical representation of the flow of execution of this topology.

It starts by receiving in the GPS spout the data about outdoors mobility. This information is forwarded by the web service as single GPS coordinates reads and is in the queue ready to be dealt by the topology. From this point on and throughout all the topology the stream of data is grouped

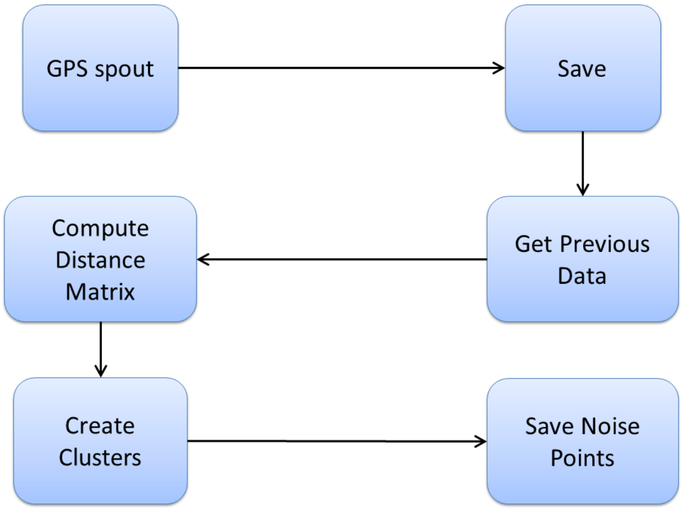


Figure 5.2: Graphical representation of the flow of execution of the GPS topology

by user, in order to guarantee that the information of each user is always treated by the same task. From the spout the GPS read data is sent to the Save bolt that persists all the information on to the Cassandra database. After storage of the received read past movement information is fetched in the Get Previous Data bolt.

Now that all the information is available the following bolts implement the DBSCAN algorithm, presented in Section 4.1.2, broken into two steps. In the first bolt, the Compute Distance Matrix, is computed the distance matrix and at the same time created a list of neighbours. This is done by taking into consideration the variable Eps that defines a radius where the neighbours of a point are. So the distance of a point to every other point is calculated and if the distance is inferior to Eps then the points are considered to be neighbours. Having the neighbour list passed to the next bolt, the Create Clusters, the clusters are created taking into consideration the other variable of the DBSCAN algorithm, the MinPts, that is the minimal number of neighbours that a point needs to have in order to be a core point of cluster. After the execution of this bolt all the clusters are discovered but what really matters are the points that are not attributed to any cluster, the noise points. These points are passed to the next and final bolt where they are persisted into the database.

Inside Movement Topology

The aim of this topology is to make the analysis of data of movement inside home. The analysis consists in checking the number of sensor reads per hour and then check if those reads comply with the normal pattern of movement of each hour established previously or not. To achieve this is used a simple statistical analysis relying in averages and standard deviation combined with the use of Dynamic Time Warping(DTW), presented in Section 4.1.7. The steps followed in this topology are represented in Figure 5.3.

The first two steps are similar to the topology presented in Figure 5.2, where the spout fetches the data from the queue and sends it to a bolt that immediately persists the information to the

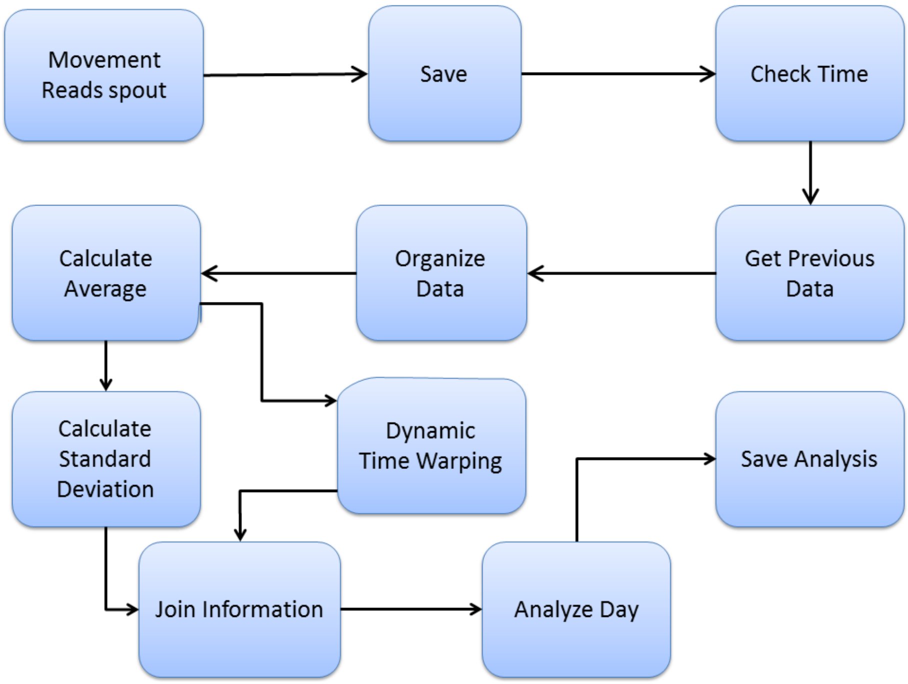


Figure 5.3: Graphical representation of the flow of execution of the Inside Movement topology

database. Then the next bolt checks the time of the read that was received. In this topology the analysis is only made when the read of the last hour of the day is received in order to make the analysis of the full day. If the read is representative of the last hour of the day then the execution continues to the bolt that will retrieve all the data collected through the last month. The execution only passes this point if there is at least a week of previous information to work with, as this information will be used to uncover the normal movement pattern, so a representative amount of information is needed. After gathering all the information this is then passed to the following bolt that is responsible to organize it in order to facilitate the analysis process that follows.

It is from this point on that the analysis of the data is done, starting by calculating the average number of reads for each hour. This allows us to establish a normal pattern of inside movement per hour. After this step the information is sent in parallel to two different bolts, one that will calculate the standard deviation and the other responsible to implement the DTW algorithm. In the DTW bolt, every hour of the day being analysed is compared with the normal pattern established, using a window of two hours what means that each hour is not only compared to the same hour of the normal pattern but also with the two previous and following ones, saving the minimum distance found. The information of the two bolts is then joined in order to be sent for the bolt responsible to do the final analysis of the day. This consists in comparing the distances found in the DTW bolt and see if they are less than two standard deviations. If they are bigger this will mean they are outside of the 95% of the values and therefore are possible warning signs. These warnings are passed to the final bolt, responsible to store the data gathered in the analysis in the database in order to be available to later presentation in the web visualization.

Lights and Leaving Home Topology

In this topology is done the implementation of two usage scenarios presented in Section 5.1.2, the number of times a person leaves home and the check of the number of lights left on when the person leaves. In Figure 5.4 is possible to see the flow of execution of this simple topology.

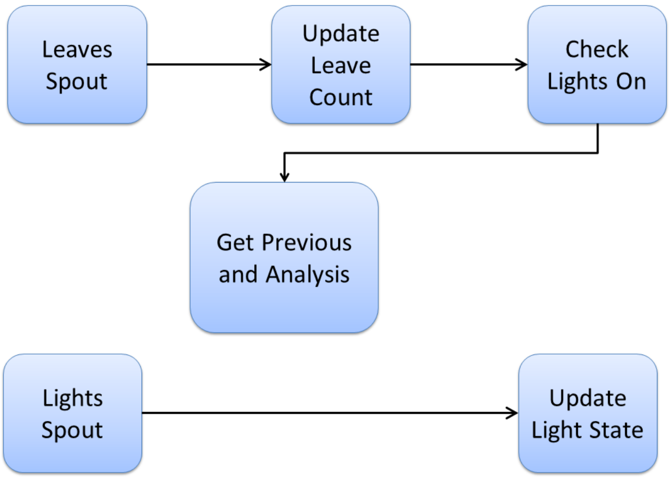


Figure 5.4: Graphical representation of the flow of execution of the Lights and Leaves topology

This topology differs from the others presented as it contains two spouts in the flow of execution. This happens because the data that comes from the two different spouts is closely related. The Lights Spout receives all the information about the state of the lights and at home sends it to the the following bolt to update that information in the database to guarantee that the correct information is present at all times. The Leaves Spout receives information every time a person leaves home. It then sends this information to the next bolt that increases the counter in the database that stores the number of leaves for that week. After updating that information the execution continues to the next bolt where it checks what lights are on at that time. Is in this step that the importance of having the update of lights running in parallel becomes evident, and by having both spouts in the same topology we guarantee that they always run closely with each other. That information is forwarded to the next bolt that retrieves from the database previous information about the number of lights left on in previous leaves and with that information checks if the number of lights left on at this time is in line with past behaviour.

Frequent Sequential Patterns Topology

The development of this topology had the objective of discovering frequent sequential patterns in data of home sensors, the Sequential Pattern Mining talked about in Section 4.1.8. In this case the aim is to uncover patterns in the behaviour of a person at home like, if a person opens the fridge usually the following step is to open the water tap. By uncovering these patterns, and possible changes on these patterns through time, possible difficulties felt by the user can become more clear, evidencing some coping techniques that the person may be using. The flow of execution of this simple topology is presented in Figure 5.5.

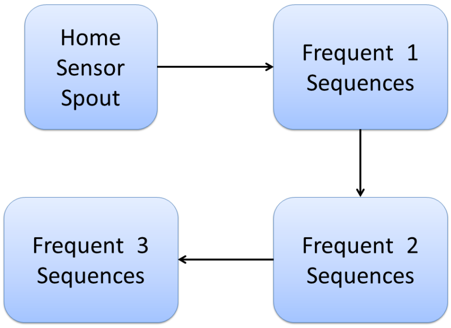


Figure 5.5: Graphical representation of the flow of execution of the Frequent Sequential Patterns topology

The entry point of data on this topology is, like in every other, the spout that reads from the correspondent queue. The message is then sent to the first bolt where is kept in memory the count of each occurrence, for example open the fridge door, for each user. Here is where the grouping of the stream of data by user, explained previously in this section, becomes important. Grouping the stream by user guarantees that all the information from one user is always process by the same task and therefore is possible to guarantee that the count kept in memory is correct. Throughout all the topology the stream is always grouped by user for this reason. This count serves to uncover frequent occurrences, that are the ones that happen more times than a threshold specified. To the next bolt is sent a list of the frequent occurrences or sequences of one element. In the second bolt is made the same process but now for sequences of two occurrences that contain one of the frequent sequences of one discovered before. The frequent sequences of two are persisted to the database as they may already offer some important insight. From this bolt is forwarded to the final bolt a list of the frequent sequences of two elements in order to make the same process but now discover frequent sequences of three elements, that are also persisted in the database. The process carried out in these bolts is known as GSP [SA96]. To uncover sequences of bigger lengths it would only be needed to add more bolts to fulfil this requirement.

5.2.3 Cassandra Database

The Cassandra database is where all the data collected and the data resultant of all the analysis is stored and therefore is a vital part of the system developed. Several reasons supported the choice to have the database using Apache Cassandra [Fou], a NoSQL and distributed database. One of the reasons is the linear scalability and fault-tolerance presented by this database. This system must be developed to handle an exponential number of small pieces of information that come in continue streams, and all that data must be continuously stored. Cassandra, running in clusters of machines, offers the possibility of replicating the information across the cluster guaranteeing that if a machine fails it will be possible to dot the read or write in some other machine. Furthermore

Cassandra displays good performances both in read and write intensive environments that will be the case of this system when deployed into production environment. [TB11]
Cassandra has a data model that differs from a traditional SQL database. The first layer of the Cassandra data model is the cluster, that is the composed by all the nodes (machines) of a Cassandra instance. A cluster may contain many keyspaces. A keyspace is the namespace that contains the column families, usually exists one keyspace per application. Column families, which can be seen as the table of a traditional SQL database, are composed by rows of columns. A row is a group of columns and is identified by an unique identifier. A column is the basic increment of data and consists of a tuple of information composed by name, value and timestamp. To further understand this structure is possible to see a graphical representation of it in Figure 5.6.

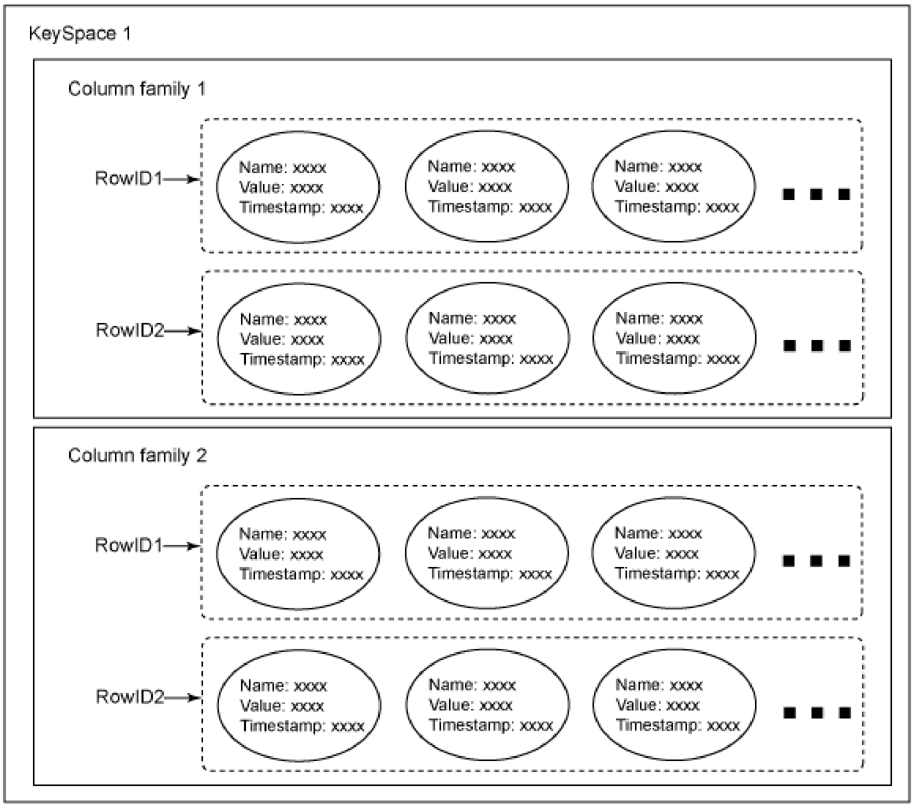


Figure 5.6: Representation of Cassandra data model

In this project were developed the following column families, contained in one keyspace, to store all the data gathered and produced by its analysis:

- **GPSRead:**
  - Where all the single reads of GPS that come to the system are stored;
  - Each row contains the columns: GPS read identifier(row key), user identifier, latitude, longitude and timestamp;

- **NoisePoint:**
  - Stores the noise points discovered by the DBSCAN execution;
  - Each row contains a GPS read identifier(row key) and user identifier;
- **MovementRead:**
  - Stores the reads of number of movements per hour inside home;
  - Each row contains the columns: read identifier(row key), user identifier, number of reads and timestamp;
- **MRAnalysedDay:**
  - Contains all the information resultant of the analysis of a day made in the Inside Movement Topology;
  - Each row contains the columns: analysed day identifier(row key), user identifier, date, averages and warnings (the hours where there is a significant deviation);
- **Light:**
  - Where the state (on or off) of the lights is stored and updated.
  - Each row contains the columns: light identifier(row key), user identifier and state;
- **NRLightsON:**
  - Information about the number of lights left on when a person leaves home;
  - Each row contains the columns: lights on identifier(row key), user identifier, number of lights on, timestamp and warning (if the number of lights on is significantly deviated from the normal);
- **LeavingCounter:**
  - Counter column family responsible to store the number of leaves per week of a user;
  - Each row contains a counter identifier(row key) and count column;
- **FrequentSeq:**
  - Stores all the frequent sequential patterns of a user;
  - Each row contains the columns: sequence identifier(row key), user identifier and sequence;

In all the column families the user identifier column is indexed to allow queries by user, as the queries must be done by indexed columns, otherwise wont work. This was done because all of the queries to the database will always be done by user, being suffice to index the user identifier as any query that has other fields will automatically work if first queried by user.

5.2.4 Web Visualization

The Web Visualization is the only part of this system that will be visible by the users and therefore the only part where the results of all the analysis made can be seen. This makes the visualization the a vital part of the whole system when viewed from the user perspective. As this is a system aimed to aid in the diagnosis and monitoring of several types of dementia there are some concerns that must be kept in mind. Is important that the information is available not only to caregivers and health professionals but also to the elders in order to be perceived by them what information is being collected by the system and what results are being produced. This further highlights the importance of having simple and clean visualizations, that shows all the different types of users the information in a clear and meaningful way, avoiding distress or unnecessary efforts to understand what is presented.
The following list presents some detailed informations of the visualizations developed:

- **Outside Movement :** In order to show the users all the information collected about outside movement and the results of its analysis was used the Version 3 of the Google Maps JavaScript API [Goob], the latest version at the time of writing. Using this it was possible to present all the information in a real map in order to allow users to relate the data with real places that are familiar to them. Three different map visualizations were developed to represent this information:
  - The GPS reads view that shows the points representing the data that was collected by the GPS tracking system. Points that were considered as outliers during the analysis are painted in red contrary to normal points that are blue.
  - The Heat Map view that shows a heat map representation of the places visited by the elder.
  - The GPS reads over Heat Map that combines the two previously referred visualizations.
- **Inside Movement :** In this visualization was aimed to show a representation of the number of registered movements inside home a person has done per hour, in a histogram like chart, highlighting the outliers found in the analysis of this data. A bar represents the number o reads for each hour and there is a line that crosses the whole graphic that represents the averages. The bars in blue represent normal reads and those in red represent the ones that deviated significantly from the normal pattern, represented by the averages line. To develop this visualization it was used the Google Chart API [Gooa].
- **Lights Left On :** This visualization, also developed with the Google Chart API [Gooa], shows to the user a representation of the number of lights left on in a user defined period of time. To achieve this an histogram of the number of lights left on when the person leaves home is created highlighting in red the ones that may represent some cause for concern, as they were considered outliers during the analysis process.

- **Number of Leaves :** To show to the users the evolution of the number of times that the elder left home per week this simple visualization, developed with the Google Chart API [Gooa], was created. It is presented to the users a line chart that shows the number of leaves of five weeks, the one selected by the user and the four previous ones. This way is possible to understand if a person is having a normal behaviour or may be leaving home less, possibly revealing signs of social withdrawal.

- **Frequent Patterns :** The objective of this visualization is to show the frequent patterns uncovered from the sensor data in order to possibly discover some abnormal combination of steps that may indicate the use of coping strategies to overcome some difficulties or even represent signs of forgetfulness. It is presented a tree that has in its root the user and then all the frequent patterns, of sequences of two or three steps, discovered for that user. This visualization was developed using D3.js [Bos]

All the visualizations can be seen in Appendix B. Furthermore, analysed examples of all the visualizations referred here will be presented in the Chapter 6 as part of the tests, results and discussion.

5.3 Contribution

As it was presented in Chapter 3 there are already solutions focused in aiding the diagnosis of some diseases or conditions, and there are even solutions focused in helping in the diagnosis and monitoring of dementia. But, as it was presented, a gap still exists as there is no solution focused on dementia that collects data from different types of sensors to track both outside movement patterns and inside home behaviours, in order to allow an early and accurate diagnosis as well as monitoring already diagnosed dementia sufferers to better follow the evolution of the cognitive problems. To do this a solution needs to establish the normal patterns of behaviour of a person and also detect possible outliers that may arise in those behaviours in order to highlight those points to make them clear to health professionals, family members and caregivers as points that are odd and a possibly alarming change in the behaviour of the elder is occurring.

This system comes to fill this gap, analysing data from several distinct sources and establishing the normal patterns of behaviour of the user both inside and outside the house. Having established the normal patterns of a user the system tries to find anomalies, i.e. possible signs of dementia, in the data that is continuously coming to the system. These anomalies, the outliers, are then highlighted in the web visualizations that make available all the information, in a simple and objective way, to health professionals, caregivers, family members and elders. This information may help in the diagnosis process, in monitoring the evolution of dementia and also give some piece of mind to the elder and to the ones responsible for him.

It is the fact of having all these functionalities put together that differentiates this project from the ones already existent, where we can see these functionalities in some of them, what further adds to their relevance, but where none has all of them combined.

Chapter 6

Tests and Evaluation

After developing such complex system some tests needed to be done in order to check if everything was working properly. The first thing to check was if all the pieces worked together as they should and then verify if the results produced during the analysis were the ones expected. Furthermore it was essential to understand if the results were relevant in terms of further understanding and easily detect the dementia symptoms described throughout Chapter 2. The system worked properly as a whole and some interesting and relevant results, presented in Section 6.1, were found.

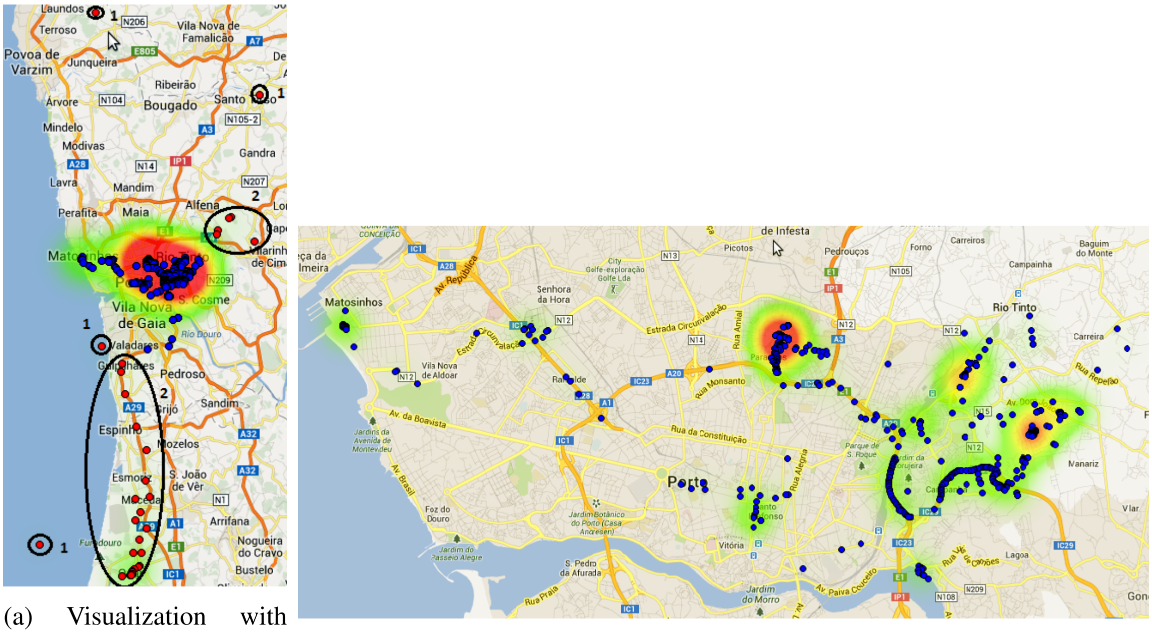
6.1 Tests and Results

Testing the system developed proved itself a challenge within the challenge of developing it. It was known from start that would be unlikely to have a smart home environment and tracking devices, collecting real data from real users, that ideally would be elders and possible dementia sufferers. To overcome this it was needed to find other possible ways of testing the whole system, particularly the analysis process. The first and more obvious solution was to find datasets that had real data, similar to the one that would be collected by the system. Unfortunately it was only possible to find datasets for two of the use cases, two GPS coordinates datasets to validate the outside movement analysis and a dataset of elders completing their normal morning routine that was used to validate the frequent sequential patterns analysis. For the scenarios were developed small Java programs to generate data that was used to test the analysis process.

6.1.1 Outside Movement tests

The aim of this analysis was to discover possible outliers in the outside movement of an elder. To do this it was implemented a DBSCAN algorithm, as stated in Section 5.2.2, to find points that were outside of the places that a person commonly visits. This analysis was made on two datasets of GPS coordinates collected using smartphones with GPS capabilities as tracking devices. In one of the cases the device worked only as a data collection platform, on the other one the device is intended to be used not only for tracking but as a navigation system as well when needed.
The data from the dataset was read from the files and sent to the system using the REST web

service presented in Section 5.2.1. It then followed the flow already presented, passing through the queues and then being read by the corresponding spout to start the analysis process done in the Storm topology, persisting all the results in the Cassandra database. After the whole process was completed it was possible to see the results presented in the graphical visualizations.



(a) Visualization with highlighted warnings

(b) Zoom in on the zone with no warnings

Figure 6.1: Outside Movement visualizations

In Figure B.1 it is possible to see the results obtained from the analysis of outside movement tracking of one user being presented in a graphical visualization with the points correspondent to GPS coordinates overlapping an heat map that represents the concentration of the reads obtained. Figure 6.1a shows the visualization that is prompted to the user when he checks his information. Here is possible to see some group of points, painted in red and circled, that are clearly deviated from the cluster of normal locations found for the user. These points were detected as outliers during the analysis process done with the DBSCAN algorithm and therefore are highlighted. These outliers appear for different reasons. The 4 points individually circled accompanied with the number 1 are probably resultant from faulty reads of the device. This faulty reads may be caused by being in indoors locations or having poor signal strength. The other points, grouped in two clusters and referred with the number 2, are the really interesting outliers in this case. These points reveal that the person moved out of the places that he usually visits and therefore a cause for concern may arise. Looking closely at the map we may assume that the group of points on the bottom may be due to vacation. In such a case the health professional and caregivers must be informed of this travel in order to disregard the information. But this points can also have other explanations and therefore it is important to have them highlighted. The other cluster thought is somewhat close to the normal locations of the user, but is not one of the common places visited by the elder. This may indicate disorientation, or a visit to a place the elder has no reason to be in, what is a cause

for concern and therefore the importance of having such points highlighted.
In Figure 6.1b is possible to see the same information but with the common locations of the user zoomed in. What seemed like a single large cluster are in reality several smaller clusters, close to each other, that are the places usually visited by the user. It is also possible to see some points that do not belong to any of those clusters but are not considered as outliers as they are close to points that do belong, and therefore do not represent cause for concern.

6.1.2 Inside Movement tests

About the movement inside home the objective was to uncover the normal pattern of movement of a person and compare everyday to the pattern established in order to understand if some change occurs. The main focus was to detect possible occurrences of night wandering, common with some cases of dementia. But this analysis may be helpful to detect other abnormalities like loss of ability to move, if the person starts to have a significant loss of movement inside its home.
To test this, as it was not possible to obtain any dataset representative of an elder movement inside home through a substantial period of time, it was created a simple Java program that produced simulated movement data. It is produced for each hour a random with, normal distribution, that represents the number of reads for each hour. In night time the reads are zero or close to zero, then there are hours of more movement in the morning, lunch and dinner time and the other hours come somewhere in the middle. The random has a normal distribution as the normal behaviour of person will also tend to have this kind of behaviour. To simulate the appearance of night wandering there are some periods of time where it may appear more movement during the night time. All this data, after being produced by the program, follows the normal path being submitted to the REST web service and then forwarded to the queues to be read by the Storm topology.
In Figure 6.2 is possible to see the results of the execution of these steps where there is no abnormal movements, all the reads comply with the normal pattern of inside movement uncovered for this user. The normal pattern is represented by the green line that crosses the whole chart, and the reads for each hour are represented by the bars. Is possible to see that the bars and the line have similar positions for each hour.

Figure 6.3 presents a case where abnormal movement was found. After the analysis of the data it is possible to see that warnings were found in hours 1, 3 and 16, painted in red and signalled in the figure. In hours 1 and 3 is clear that the number of reads are abnormal when compared with the normal pattern of movement, as there is a lot more movement then what would be expected for those hours. The warning in hour 16 is not that obvious but is explained by the value of standard deviation for that hour. The hours of the middle of the afternoon vary in fairly small amounts, having a small value of standard deviation, and small deviances may have a distance of more than two standard deviations from the normal pattern therefore generating warnings that in truth may not be cause for concern.

In spite of the possibly false positive found, the system behaved as expected as it is preferable to detect false positives than missing some real abnormal values that should be highlighted as

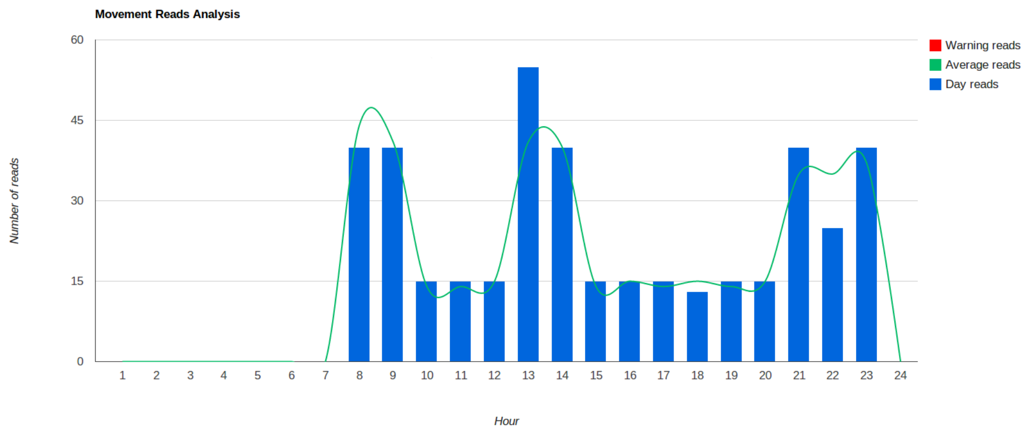


Figure 6.2: Representation of inside movement with no warnings

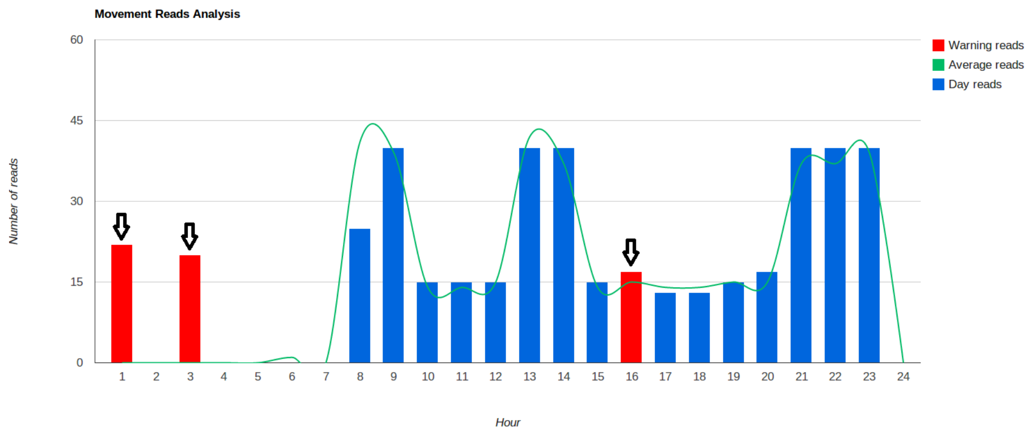


Figure 6.3: Representation of inside movement with warnings

warnings. Furthermore, the registers that might indicate night wandering were clearly identified, and that was the major objective in this analysis.

6.1.3 Lights On and Leaving Home tests

In this case there are two different objectives that are aimed, the detection of signs of forgetfulness, in this case forgetting to leave the lights off when leaving home and monitor the number of times a person leaves home in order to be perceivable if some signs of social withdrawal start to appear. This is the other case where real world datasets were not found, and therefore the approach of creating a small Java program to produce simulated data was followed. It was produced simulated data for a period of three months where a person would leave home a random number of times per day and at the time of leaving data for five lights inside home, identified as bathroom, bedroom, kitchen, living room and hallway, was updated. In this way it was being sent in parallel to the system the update of the lights states and the information that a person was leaving home with a time stamp. The information followed the same process path as the presented previously in order to test the system as a whole and having the information available for the analysis.

Figures 6.4 and 6.5 present the results of the analysis done on the data produced. In 6.4 the data presented covers one entire month while in 6.5 is possible to see the data from only two weeks. This allows to look to the data with different with different granularities what may allow different insights. Both charts presented highlighted are, bars that are red and pointed out, that are the ones that were considered as warning signs after the analysis of the data.
It is also possible to notice that in the right end of both charts there are bars representing high numbers of lights left on, but not considered as warnings. This happens because past data is used to calculate what is the normal number of lights left on by a user. In this case, because a lot of lights have been left on in the past few leaves this stopped to be considered as outlier and was dismissed as a concept drift. This is exactly what was expected because when there was a change in behaviour this was detected but when that behaviour started to be the normal one it is essential to understand that a concept drift occur and no longer makes sense to differentiate that data.
In Figure 6.6 it is possible to see a chart representing the evolution of the number of times a person left home per week in the past five weeks. The information presented in this chart is simply the count of the leaves but is easy to see when some drastic decrease in the number of leaves occurs this will be easily seen, allowing to detect a possible case of social withdrawal, a serious problem with dementia sufferers.

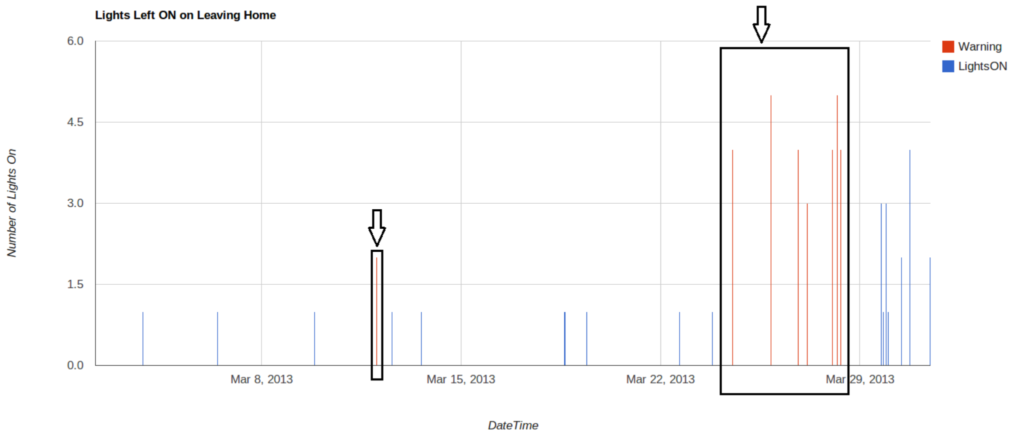


Figure 6.4: Chart with the number of lights left on when leaving home through the course of one month

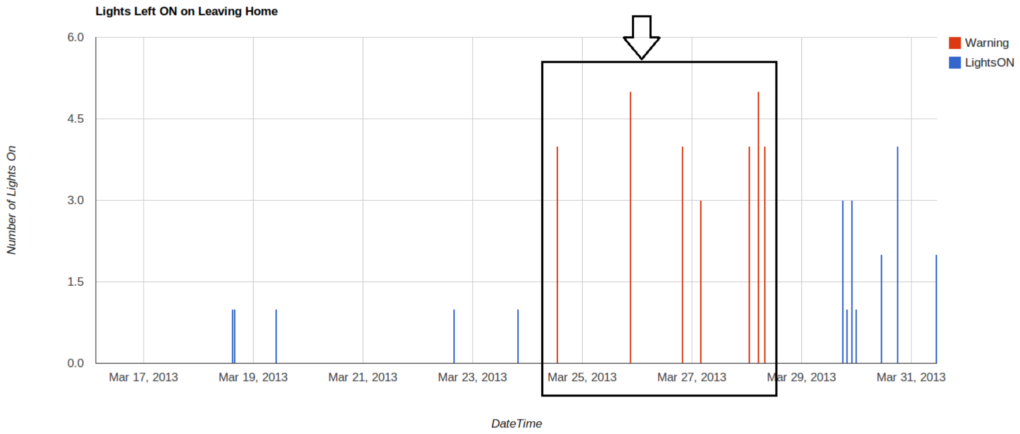


Figure 6.5: Chart with the number of lights left on when leaving home through the course of two weeks

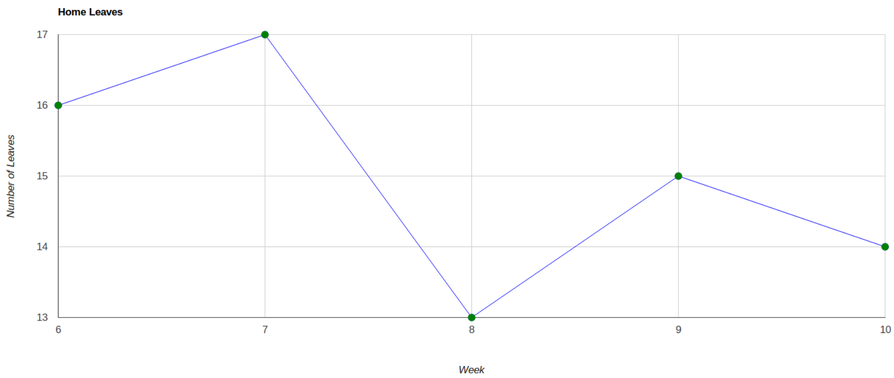


Figure 6.6: Leaving home count per week

6.1.4 Frequent Sequential Patterns tests

The final test to be presented is the uncovering of frequent sequential patterns from a user. To test this functionality it was used a dataset where six elders did their morning routines in a smart home environment that collected data from several sensors in doors, lights, taps and other appliances. This data was read from the dataset files and sent to the REST web service following the same path as every other piece of information in order to reach the analysis process.
Figure 6.7 represents the frequent patterns found for one of the users. It is possible to see in the representation both sequences of two elements, like opening and closing the cold water tap, and of three elements, like turn on the toaster, turning it of and then opening the kitchen sink. This graphical representation of the data can give us some insights of the steps an elder takes to do some tasks, and knowing those steps can allow us to make inferences of some of the difficulties the person might be feeling. If a person repeats some steps many times that may be a possible sign of forgetfulness.
Furthermore this data, when looked at in different periods of time can show us if some new frequent patterns arise and that patterns can offer some possible insights into the evolution of cogni-

tive impairments. As the data used was only collect in a few days this option was not explored, because such short span of time would not allow to do this kind of analysis.
Also in this case the system worked as it was pretended, with the information following the correct path through the system, and with the analysis producing the kind of results that would be desirable, representing the frequent sequential patterns of each user.

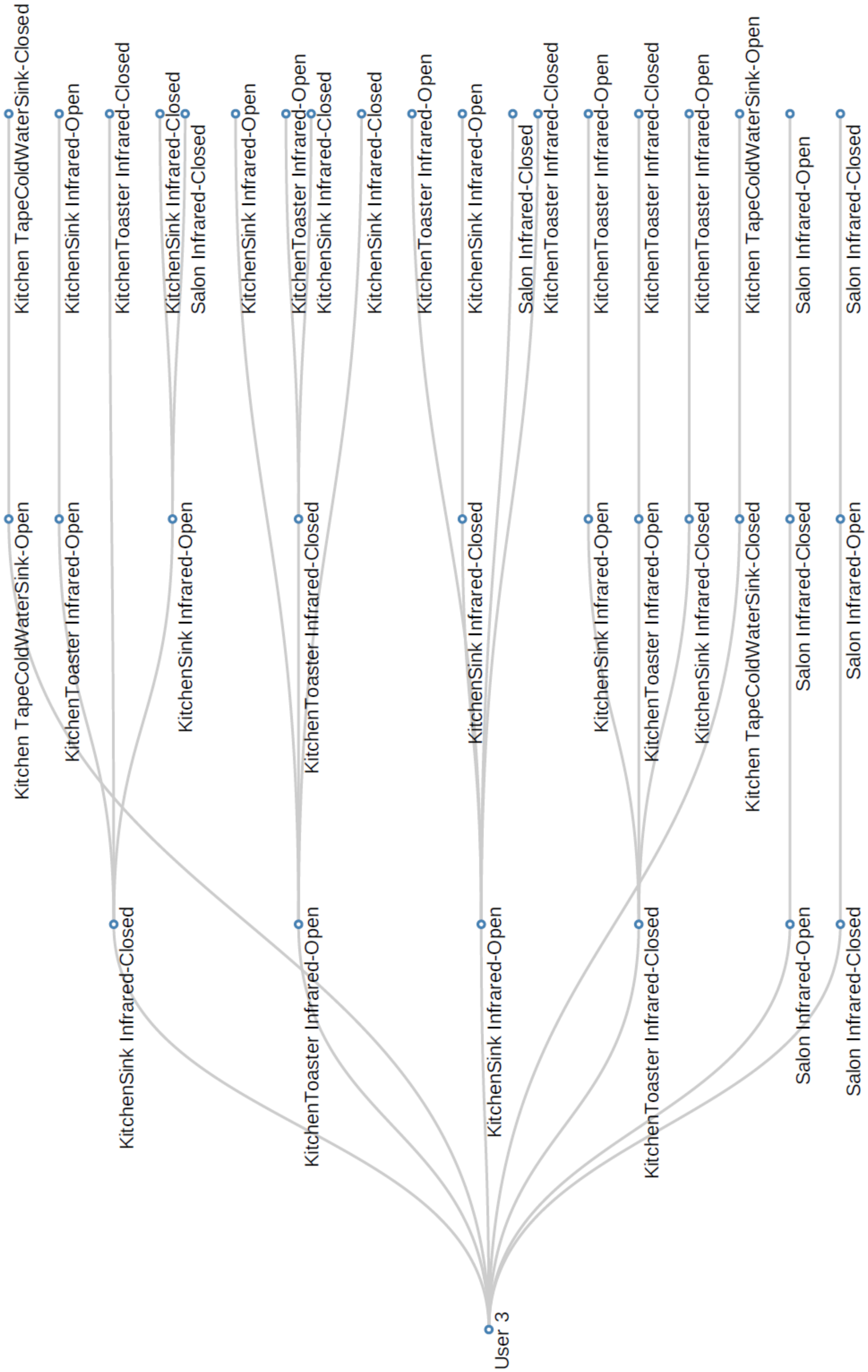


Figure 6.7: Graphical view of the frequent patterns found for a user

6.2 Results Discussion

The results presented in Section 6.1 allow us to infer that the system presented could prove itself a valuable aid in the monitoring and diagnosis of different types of dementia by collecting different types of data and highlighting possible warning signs. There are obviously some questions that would need to be addressed in order to make this product ready to be deployed in a production environment. First there is the question of testing all the implemented features with real data, in order to further validate them. Then, it is also necessary to design sensing platforms such as a smart home environments in order to have a really complete system. Despite this, the system was able to perform the tasks that was aimed to do and as it is possible to have different types of data added it could be easily included in other solutions already available, making only few small adjustments in order to correctly treat and analyse the concrete types of information produced by those systems.

First, from the several tests run it was possible to establish that the system as a whole worked without problems, with all the parts developed working together and communicating with each other with no problems. The REST web service responded correctly to every request, both of input of data and the gathering of information for the visualizations. The Storm topologies also worked as expected, analysing the streams of data produced and sent to the system in a timely fashion, making the analysis of the data in real-time as it was aimed. The Cassandra database worked as pretended not causing delays in the execution of the topologies, allowing to maintain all the responses in real time. Obviously further testing, with streams of real data being produced in real time, is needed to better support the good results obtained. Also it will be important to deploy the Storm topologies and the Cassandra database on to real clusters to test the system for the amounts of data that would be generated in a production environment.

All the different analysis made produced relevant data, that may be helpful to dementia sufferers and possible undiagnosed people with dementia. To further validate the value of this data it would be important to have a questionnaire addressed to caregivers, family members, health professionals and people with dementia. Even so, based on the knowledge gathered in bibliography and presented in Chapter 2, we believe that the results obtained are promising.

The outside movement analysis produced interesting and valid data, uncovering points that, when related with all the GPS information gathered, seem odd as they are too further apart from the others and may indicate visits to places that are not part of the normal routine of the elder, which may indicate disorientation.

The results obtained from the analysis of the inside movement seemed to show that the system is able to detect significant changes in the normal patterns of a user. With the tests made was possible to see that the analysis process was able to detect possible cases of night wandering, common in persons with dementia.

In the analysis of the number of lights left on were detected the cases that may show signs of forgetfulness and it was also possible to track the evolution of the number of times a person left home per week, a factor that may give insight to possible signs of social withdrawal.

Finally from the analysis the frequent sequential pattern analysis found the patterns that are common for a user, allowing to understand how a person tackles some of the tasks of its everyday living. More than that the analysis of the evolution of those patterns, what new patterns arise and what patterns stop to be common, may allow to understand if some impairment is occurring.
In summary, the results obtained seem promising, showing that this system, when allied with a proper sensing platform composed by smart home environment with sensors, GPS tracking and other sources of information may in fact prove to be a valuable ally to health professionals in allowing an earlier and correct diagnose and also help to follow the evolution of the disease in already diagnosed elders. Furthermore this system may help family members, caregivers and elders themselves by giving them more information, in an understandable way, possibly allowing some more peace of mind and freedom to all.

Chapter 7

Conclusions and Future Work

In this project it was designed and implemented a system that relies in several technologies oriented for the future, distributing the workload in order to be able to tackle large amounts of real data being continuously produced in a streaming fashion, making possible to create a solution to aid and assist in the problem of diagnosing and monitoring persons with dementia. This is a problem of the present and, as data shows, it may become an even more serious problem in the future. The usage of these new technologies, in particular Storm and Cassandra, made the project very interesting from a developer point of view. Using such technologies allowed to be in touch with new concepts and ways of dealing with problems in a different way. This solutions are oriented to analyse great amounts of data and make the results available in real time, making them ideal to treat streams of data and make the results available in a timely manner. As it was expected the usage of such technologies posed some challenges as the documentation and solutions for them are sometimes difficult to find and the rationale behind the solutions is sometimes different, but the benefit of exploring new solutions that may prove to be the future in data analysis and storage, and are already the present for some companies, made the project even more appealing.
Furthermore being able to study some different data mining techniques and algorithms, implementing some of them in a distributed way as part of the solution developed, allowed to gain deeper knowledge in an vast and interesting field, which can embedded in systems of several areas from health care, as in the case presented, to economy or advertisement not mentioning many others.

7.1 Objectives Fulfilment

The main goal of this dissertation work was to design and implement a system capable of detecting possible signs of dementia using monitoring data from the daily activities of an elder. This dissertation aimed to use new technologies and data mining techniques, in particular outlier detection, to achieve its goal and therefore be a valuable and valid helper in the diagnosis and monitoring of different types of dementia. In order to achieve these goals some objectives were defined. After the development and testing this dissertation produced a functional proof of concept prototype that accomplishes the proposed objectives, as it is capable of:

- Making real-time analysis of streams of data;
- Being easily extendible, it is only necessary to develop new topologies to analyse new types of data and submit them to the running cluster, having no interference with the analysis already implemented;
- Analysing different types of information, both from outside movement and inside behaviours;
- Making all the analysis process seamless to the user, not needing configuration or other inputs;
- Uncover the normal behaviour pattern of the user and finding outliers in data that may be indicating signs of dementia;
- Presenting the information collected and resultant from the analysis in a simple and clear way.

To make this system complete it would be needed a suited sensing platform, not available during the development and that was not possible to develop during the dissertation. In spite of this, as the primary objective was to create the data analysis part of the system, all the main objectives were fulfilled during the development of the dissertation.
In order to verify that all the functionalities implemented fulfilled the primary objective of detecting possible early signs of dementia the system was tested with different types of data. After running all this data through the system, analysing and storing it, it was possible to see that these signs were indeed detected. Using GPS data it was possible to find outliers that may indicate disorientation as they were points that did not belong to the normal day to day living of the user. With inside home movement data it was possible to detect possible cases of night wandering by establishing the normal movement pattern of an user and then try to find deviations in the data that was arriving. It was also possible to detect signs of forgetfulness by checking the number of lights left on by the user when leaving home. By monitoring the number of times a person leaves home it was also possible to perceive possible indications of social withdrawal. And using data from everyday living tasks is possible to uncover patterns of steps taken by the users to complete these tasks, what allows to see how this tasks are being carried out. We believe that all this signs detected are valuable in the diagnosis and monitoring process of persons with dementia, making this system a valuable aid in these processes.

7.2 Future Work

A system such as the one described throughout this dissertation is complex with several parts that need to be carefully developed and tested, making the work developed only a basis and starting

work for a system that ultimately will be composed by other parts.
First of all there is obviously room for improvement in the functionalities already implemented. Particularly in the frequent sequential patterns it would be important to allow the user to have in the graphical visualization a time-line that would allow the user to easily see the frequent patterns for different periods of time in order to easily understand possible evolutions that may occur. The graphical visualizations in general also need to be further improved adding further information and functionalities. The visualizations need to be validated through thorough usability testing, especially with elder users, as they may be the ones feeling more difficulties using this tool. It is also important to create a general interface with a simple user page where the user can choose to view the different types of data collected and its analysis.
Adding to that there were several lines of work not explored during the development of the dissertation but that need to be addressed in order to make this a complete system, ready for production. Furthermore this work would allow the system to be more be valuable as a diagnosis and monitoring tool for people with dementia or who present some kind of cognitive impairments that may represent the early signs of dementia.
The main feature that lacks in this system is the sensing platforms essential to collect all the information that after analysis may be useful to the diagnosis process. These sensing platforms consist in home sensors capable of collecting data about movement inside house as well as sensors installed in all kinds of appliances and home utilities in order to have data to understand the common patterns of one person. These sensors must also be able to know when the lights are on or off and also register when the person leaves home in order to be capable of analysing possible signs of forgetfulness like leaving lights on when leaving home.
Also in the analysis part of the system, the one developed throughout the dissertation some new features need to be added in order to make this a complete system. Some of those features are presented in Section 5.1.2 like detect differences in the paths usually taken by the user that may be indicating of disorientation but are not much different from the places usually visited by the elder or monitor the taking of medication in order to capture possible signs of forgetfulness. Other analysis features might be uncovered by questioning health professionals, caregivers, family members and elders in order to be possible to understand what functionalities they would deem useful in such a system.
So finally another possible future work that would improve the system would be the creation of a questionnaire where it would be possible to understand what functionalities would be valuable to the final users of such a system.

References

[AAS10] Tariq Ali, Sohail Asghar, and Naseer Ahmed Sajid. Critical analysis of dbscan variations. Karachi, Pakistan, 2010.

[AJK10] Malak Alshawabkeh, Byunghyun Jang, and David Kaeli. Accelerating the local outlier factor algorithm on a gpu for intrusion detection systems. In *Proceedings of the 3rd Workshop on General-Purpose Computation on Graphics Processing Units*, GPGPU '10, pages 104–110, New York, NY, USA, 2010. ACM.

[ANMF10] Nicola Armstrong, Chris Nugent, George Moore, and Dewar Finlay. Using smartphones to address the needs of persons with alzheimers disease. *annals of telecommunications - annales des telecommunications*, 65(9-10):485–495, 2010.

[AP02] Fabrizio Angiulli and Clara Pizzuti. Fast outlier detection in high dimensional spaces. In *Proceedings of the 6th European Conference on Principles of Data Mining and Knowledge Discovery*, PKDD '02, pages 15–26, London, UK, UK, 2002. Springer-Verlag.

[AZL06] Naoki Abe, Bianca Zadrozny, and John Langford. Outlier detection by active learning. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '06, pages 504–509, New York, NY, USA, 2006. ACM.

[Bar78] V. Barnett. The study of outliers: purpose and model. *Applied Statistics*, 27(3):242–250, 1978.

[BFSM12] Jakob E. Bardram, Mads Frost, Károly Szántó, and Gabriela Marcu. The monarca self-assessment system: a persuasive personal monitoring system for bipolar patients. In *Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium*, IHI '12, pages 21–30, New York, NY, USA, 2012. ACM.

[BKNS99] Markus M. Breunig, Hans-Peter Kriegel, RaymondT. Ng, and Jorg Sander. Optics-of: Identifying local outliers. In JanM. Zytkow and Jan Rauch, editors, *Principles of Data Mining and Knowledge Discovery*, volume 1704 of *Lecture Notes in Computer Science*, pages 262–270. Springer Berlin Heidelberg, 1999.

[BKNS00] Markus M. Breunig, Hans-Peter Kriegel, Raymond T. Ng, and Jorg Sander. Lof: identifying density-based local outliers. In *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, SIGMOD '00, pages 93–104, New York, NY, USA, 2000. ACM.

[BLA+09] Maged N Kamel Boulos, Ricardo Castellot Lou, Athanasios Anastasiou, Chris D Nugent, Jan Alexandersson, Gottfried Zimmermann, Ulises Cortes, and Roberto

Casas. Connectivity for healthcare and well-being management: examples from six european projects. *International Journal Of Environmental Research And Public Health*, 6(7):1947 – 1971, 2009.

[Bos] Michael Bostock. D3.js - data-driven documents. http://d3js.org/. Online; Accessed June 6th, 2013.

[BRM+07] Maged N Kamel Boulos, Artur Rocha, Angelo Martins, Manuel Escriche Vicente, Armin Bolz, Robert Feld, Igor Tchoudovski, Martin Braecklein, John Nelson, Gearoid O Laighin, Claudio Sdogati, Francesca Cesaroni, Marco Antomarini, Angela Jobes, and Mark Kinirons. Caalyx: a new generation of location-based services in healthcare. *International Journal Of Health Geographics*, 6:9, 2007.

[CBK07] Varun Chandola, Arindam Banerjee, and Vipin Kumar. Outlier detection: A survey, 2007.

[CDRV09] Kaare Christensen, Gabriele Doblhammer, Roland Rau, and James W. Vaupel. Ageing populations: the challenges ahead. *Lancet*, 374(9696):1196 – 1208, 2009.

[CH06] R.J.G.B. Campello and E.R. Hruschka. A fuzzy extension of the silhouette width criterion for cluster analysis. *Fuzzy Sets and Systems*, 157(21):2858 – 2875, 2006.

[DND+09] R.J. Davies, C.D. Nugent, M.P. Donnelly, M. Hettinga, F.J. Meiland, F. Moelaert, M.D. Mulvenna, J.E. Bengtsson, D. Craig, and R.-M. Droes. A user driven approach to develop a cognitive prosthetic to address the unmet needs of people with mild dementia. *Pervasive and Mobile Computing*, 5(3):253 – 267, 2009.

[EKSX96] M. Ester, H.-P. Kriegel, J. Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. pages 226 – 31, Menlo Park, CA, USA, 1996.

[FLGGV+11] Carlos Fernandez-Llatas, Juan Miguel Garcia-Gomez, Javier Vicente, Juan Carlos Naranjo, Monserrat Robles, Jose Miguel Benedi, and Vicente Traver. Behaviour patterns detection for persuasive design in nursing homes to help dementia patients. *Conference Proceedings: ... Annual International Conference Of The IEEE Engineering In Medicine And Biology Society. IEEE Engineering In Medicine And Biology Society. Conference*, 2011:6413 – 6417, 2011.

[Fou] The Apache Software Foundation. Cassandra. http://cassandra.apache.org/. Online; Accessed June 3rd, 2013.

[Gooa] Google. Google charts. https://developers.google.com/chart/. Online; Accessed June 6th, 2013.

[Goob] Google. Google maps javascript api v3. https://developers.google.com/maps/documentation/javascript/. Online; Accessed June 6th, 2013.

[HA04] Victoria Hodge and Jim Austin. A survey of outlier detection methodologies. *Artif. Intell. Rev.*, 22(2):85–126, October 2004.

[HCK+05] V. Hautamaki, S. Cherednichenko, I. Karkkainen, T. Kinnunen, and P. Franti. Improving k-means by outlier removal. volume 3540, pages 978–987, 2005.

[HCXY07] Jiawei Han, Hong Cheng, Dong Xin, and Xifeng Yan. Frequent pattern mining: current status and future directions. *Data Min. Knowl. Discov.*, 15(1):55–86, August 2007.

[Hol12] C. Holmes. Dementia. *Medicine (United Kingdom)*, 40(11):628–631, 2012.

[HRK$^+$10] John Hicks, Nithya Ramanathan, Donnie Kim, Mohamad Monibi, Joshua Selsky, Mark Hansen, and Deborah Estrin. Andwellness: an open mobile system for activity and experience sampling. In *Wireless Health 2010*, WH '10, pages 34–43, New York, NY, USA, 2010. ACM.

[Int09] Alzheimer Disease International. World alzheimer report 2009, 2009.

[Int11] Alzheimer Disease International. World alzheimer report 2011, 2011.

[KL83] J. Kruskal and M. Liberman. The symmetric time-warping problem: from continuous to discrete. *In Time Warps, String Edits and Macromolecules: The Theory and Practice of Sequence Comparison*, pages 125–161, 1983.

[KMN$^+$02] Tapas Kanungo, D.M. Mount, N.S. Netanyahu, C.D. Piatko, R. Silverman, and A.Y. Wu. An efficient k-means clustering algorithm: analysis and implementation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 24(7):881–892, 2002.

[KShZ08] Hans-Peter Kriegel, Matthias S hubert, and Arthur Zimek. Angle-based outlier detection in high-dimensional data. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '08, pages 444–452, New York, NY, USA, 2008. ACM.

[Lab] Quantified Self Labs. Quantified self - self knowledge through numbers. http://quantifiedself.com/. Online; Accessed May 20th, 2013.

[Lee10] Matthew L. Lee. Embedded assessment of wellness with smart home sensors. In *Proceedings of the 12th ACM international conference adjunct papers on Ubiquitous computing - Adjunct*, Ubicomp '10 Adjunct, pages 473–476, New York, NY, USA, 2010. ACM.

[LWLW08] Guiling Li, Yuanzhen Wang, Min Li, and Zongda Wu. Similarity match in time series streams under dynamic time warping distance. In *Computer Science and Software Engineering, 2008 International Conference on*, volume 4, pages 399–402, 2008.

[Mac03] David J. C. Mackay. *Information Theory, Inference and Learning Algorithms*. Cambridge University Press, first edition edition, June 2003.

[MAMM12] James McNaull, Juan Carlos Augusto, Maurice Mulvenna, and Paul McCullagh. Data and information quality issues in ambient assisted living systems. *J. Data and Information Quality*, 4(1):4:1–4:15, October 2012.

[Mara] Nathan Marz. Storm - distributed and fault-tolerant realtime computation. http://storm-project.net/. Online; Accessed May 20th, 2013.

[Marb] Nathan Marz. Storm wiki. https://github.com/nathanmarz/storm/wiki. Online; Accessed June 4th, 2013.

[MC05] I. McKeith and J. Cummings. Behavioural changes and psychological symptoms in dementia disorders. *Lancet Neurology*, 4(11):735–742, 2005. cited By (since 1996) 51.

[MDH08] L.F. Mendes, Bolin Ding, and Jiawei Han. Stream sequential pattern mining with precise error bounds. In *Data Mining, 2008. ICDM '08. Eighth IEEE International Conference on*, pages 941–946, 2008.

[Mis05] F Miskelly. Electronic tracking of patients with dementia and wandering using mobile phone technology. *Age & Ageing*, 34(5):497 – 499, 2005.

[MRBK+07] F J M Meiland, A Reinersmann, B Bergvall-Kareborn, D Craig, F Moelaert, M D Mulvenna, C Nugent, T Scully, J E Bengtsson, and R M Droes. Cogknow development and evaluation of an ict-device for people with mild dementia. *Studies In Health Technology And Informatics*, 127:166 – 177, 2007.

[Naz11] Linda Nazarko. Understanding dementia: diagnosis and development. *British Journal of Healthcare Assistants*, 5(5):216 – 220, 2011.

[OI12] World Health Organization and Alzheimer Disease International. Dementia: A public health priority, April 2012.

[PWH12] AM Pot, BM Willemse, and S Horjus. A pilot study on the use of tracking technology: Feasibility, acceptability, and benefits for people in early stages of dementia and their informal caregivers. *Aging & Mental Health*, 16(1):127 – 134, 2012.

[RRS00] Sridhar Ramaswamy, Rajeev Rastogi, and Kyuseok Shim. Efficient algorithms for mining outliers from large data sets. *SIGMOD Rec.*, 29(2):427–438, May 2000.

[SA96] Ramakrishnan Srikant and Rakesh Agrawal. Mining sequential patterns: Generalizations and performance improvements. In *Proceedings of the 5th International Conference on Extending Database Technology: Advances in Database Technology*, EDBT '96, pages 3–17, London, UK, UK, 1996. Springer-Verlag.

[SAF+08] Noam Shoval, Gail K Auslander, Tim Freytag, Ruth Landau, Frank Oswald, Ulrich Seidl, Hans-Werner Wahl, Shirli Werner, and Jeremia Heinik. The use of advanced tracking technologies for the analysis of mobility in alzheimer's disease and related cognitive diseases. *BMC Geriatrics*, 8:7, 2008.

[SC07] Stan Salvador and Philip Chan. Toward accurate dynamic time warping in linear time and space. *Intell. Data Anal.*, 11(5):561–580, October 2007.

[SG11] Shiblee Sadik and Le Gruenwald. Online outlier detection for data streams. pages 88 – 96, Lisbon, Portugal, 2011. Data distribution;Data points;Data sets;Data stream;Execution time;Online data;Online implementation;Outlier Detection;Performance study;Random access;.

[Sha09] JY Shagam. The many faces of dementia. *Radiologic Technology*, 81(2):153 – 168, 2009.

[SL] Inderjeet Singh and Joel Leitch. Google gson. https://sites.google.com/site/gson/Home. Online; Accessed June 4th, 2013.

[TB11] B.G. Tudorica and C. Bucur. A comparison between several nosql databases with comments and notes. In *Roedunet International Conference (RoEduNet), 2011 10th*, pages 1–5, 2011.

[Tsy04] Alexey Tsymbal. The problem of concept drift: Definitions and related work. Technical report, Department of Computer Science, Trinity College Dublin, Ireland, 2004.

[WHOD03] S. Welsh, A. Hassiotis, G. O'mahoney, and M. Deahl. Big brother is watching you–the ethical implications of electronic surveillance measures in the elderly with dementia and in adults with learning difficulties. *Aging & Mental Health*, 7(5):372–375, 2003. PMID: 12959806.

[WKRQ+07] Xindong Wu, Vipin Kumar, J. Ross Quinlan, Joydeep Ghosh, Qiang Yang, Hiroshi Motoda, Geoffrey J. McLachlan, Angus Ng, Bing Liu, Philip S. Yu, Zhi-Hua Zhou, Michael Steinbach, David J. Hand, and Dan Steinberg. Top 10 algorithms in data mining. *Knowl. Inf. Syst.*, 14(1):1–37, December 2007.

[WPZZ12] Pang Wu, Huan-Kai Peng, Jiang Zhu, and Ying Zhang. Senscare: Semi-automatic activity summarization system for elderly care. In JoyYing Zhang, Jarek Wilkiewicz, and Ani Nahapetian, editors, *Mobile Computing, Applications, and Services*, volume 95 of *Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering*, pages 1–19. Springer Berlin Heidelberg, 2012.

REFERENCES

Appendix A

JSON responses to the visualizations

A.1 JSON Outside Movement view

Example of JSON sent to Outside Movement visualizations


```
{
   "gpsInfo":[
      {
         "lat":"33.8051408529282",
         "lon":"-118.05576145649",
         "outlier":"N"
      },
      ... //All the points are in the JSON response
      {
         "lat":"33.7769454717636",
         "lon":"-118.121389746666",
         "outlier":"Y"
      }
   ],
   "max_lat":33.9091354608536,
   "min_lat":33.6915922164917,
   "max_lon":-117.913738489151,
   "min_lon":-118.134570121765,
   "center_lat":33.80036383867265,
   "center_lon":-118.024154305458
}
```

A.2 JSON to Inside Movement View

Example of JSON sent to Inside Movement visualizations


```
{
   "table":{
      "cols":[
         {
            "id":"Hour",
            "label":"Hour",
            "type":"number"
         },
         {
            "id":"Day",
            "label":"Day reads",
            "type":"number"
         },
         {
            "id":"Average",
            "label":"Average reads",
            "type":"number"
         },
         {
            "id":"Warnings",
            "label":"Warning reads",
            "type":"number"
         }
      ],
      "rows":[
         {
            "c":[
               {
                  "v":1.0
               },
               {
                  "v":0.0
               },
               {
                  "v":0.0
               },
               {
```

```
                    "v":22.0
                }
            ]
        },
        ...
        {
            "c":[
                {
                    "v":24.0
                },
                {
                    "v":0.0
                },
                {
                    "v":0.0
                },
                {
                    "v":0.0
                }
            ]
        }
    ]
},
"warns":[
    0,
    2,
    15
]
}
```


A.3 JSON to Lights Information View

Example of JSON sent to Lights Information visualization


```
[
    {
        "tsUID":"1364245789184-1",
        "UID":"1",
        "LON":5,
        "ts":1364245789184,
```

```
        "warn":"YES"
    },
    {
        "tsUID":"1364328589184-1",
        "UID":"1",
        "LON":4,
        "ts":1364328589184,
        "warn":"YES"
    },
          ... One register for each light on Information
    {
        "tsUID":"1364728189184-1",
        "UID":"1",
        "LON":2,
        "ts":1364728189184,
        "warn":"NO"
    }
]
```


A.4 JSON to Leaving Home Information View

Example of JSON sent to Leaving Home Information visualization


```
{
    "cols":[
        {
            "id":"Week",
            "label":"Week",
            "type":"number"
        },
        {
            "id":"Leaves",
            "label":"Leaves",
            "type":"number"
        }
    ],
    "rows":[
        {
            "c":[
                {
```

```
                    "v":8.0
                },
                {
                    "v":13.0
                }
            ]
        },
        {
            "c":[
                {
                    "v":9.0
                },
                {
                    "v":15.0
                }
            ]
        },
        {
            "c":[
                {
                    "v":10.0
                },
                {
                    "v":14.0
                }
            ]
        },
        {
            "c":[
                {
                    "v":11.0
                },
                {
                    "v":11.0
                }
            ]
        },
        {
            "c":[
                {
```

```
                "v":12.0
            },
            {
                "v":16.0
            }
        ]
    }
  ]
}
```


A.5 JSON to Frequent Sequential Patterns View


```
{
   "name":"User 1",
   "children":[
      {
         "name":"Kitchen TapeHotWaterSink-Closed",
         "children":[
            {
               "name":"Kitchen TapeHotWaterSink-Open",
               "size":1
            },
            {
               "name":"Kitchen TapeColdWaterSink-Closed",
               "size":1
            }
         ]
      },
      {
         "name":"KitchenSink Infrared-Open",
         "children":[
            {
               "name":"KitchenSink Infrared-Closed",
               "children":[
                  {
                     "name":"KitchenSink Infrared-Open",
                     "size":1
                  }
               ]
```

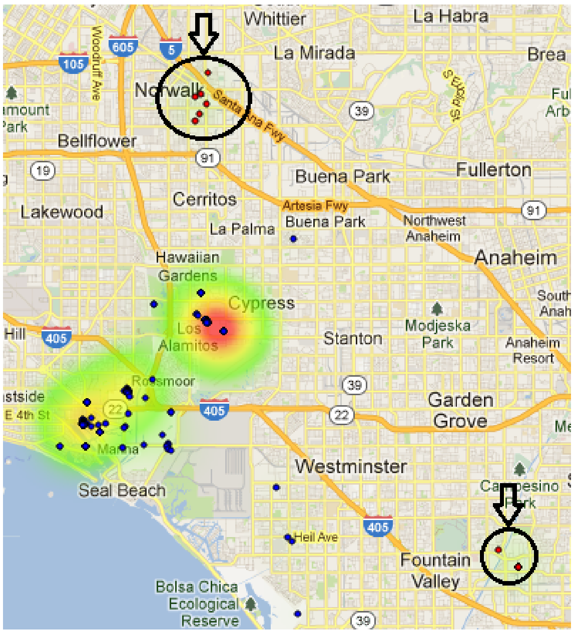
```
                }
            ]
        },
        {
            "name":"Kitchen TapeColdWaterSink-Open",
            "children":[
                {
                    "name":"Kitchen TapeColdWaterSink-Closed",
                    "size":1
                }
            ]
        },
        {
            "name":"Kitchen TapeColdWaterSink-Closed",
            "children":[
                {
                    "name":"KitchenSink Infrared-Closed",
                    "size":1
                },
                {
                    "name":"Kitchen TapeColdWaterSink-Open",
                    "size":1
                }
            ]
        },
        {
            "name":"KitchenSink Infrared-Closed",
            "children":[
                {
                    "name":"KitchenSink Infrared-Open",
                    "children":[
                        {
                            "name":"KitchenSink Infrared-Closed",
                            "size":1
                        }
                    ]
                }
            ]
        },
        {
```

```
            "name":"Kitchen TapeHotWaterSink-Open",
            "children":[
               {
                  "name":"Kitchen TapeColdWaterSink-Open",
                  "size":1
               },
               {
                  "name":"Kitchen TapeHotWaterSink-Closed",
                  "size":1
               }
            ]
         }
      ]
}
```

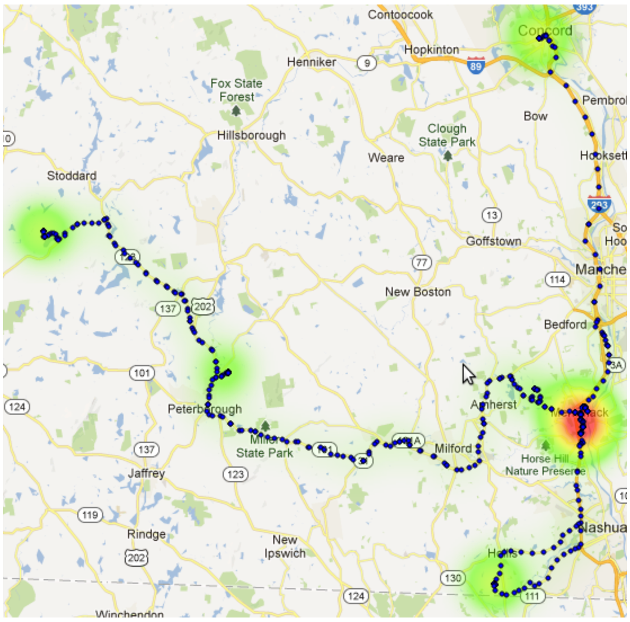
Appendix B

Web Visualizations

B.1 Outside Movement visualizations



(a) Example of outside movement visualization with warnings



(b) Example of inside movement visualization with no warnings

Figure B.1: Examples of the outside Movement visualizations

B.2 Inside Movement visualization

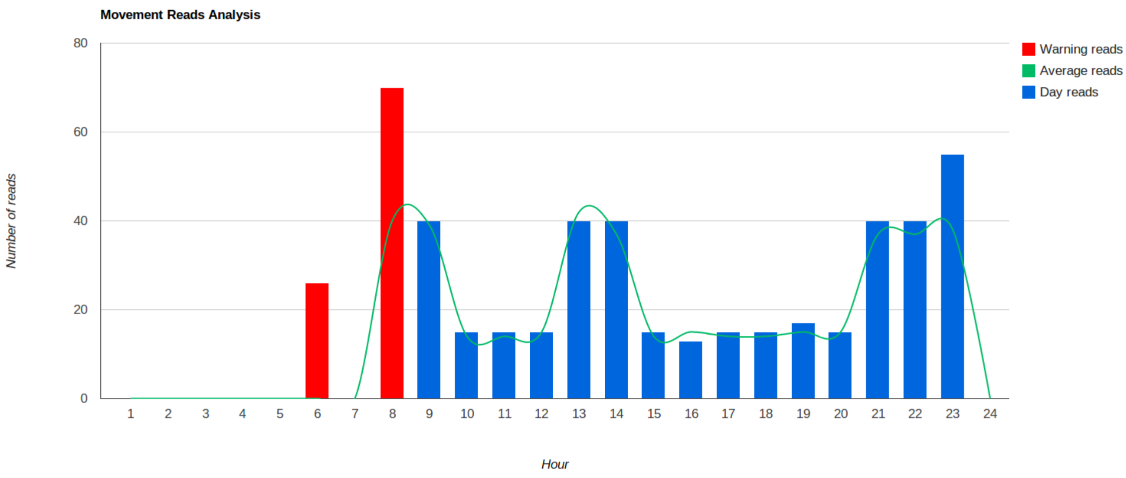


Figure B.2: Example of inside movement visualization

B.3 Lights On visualization

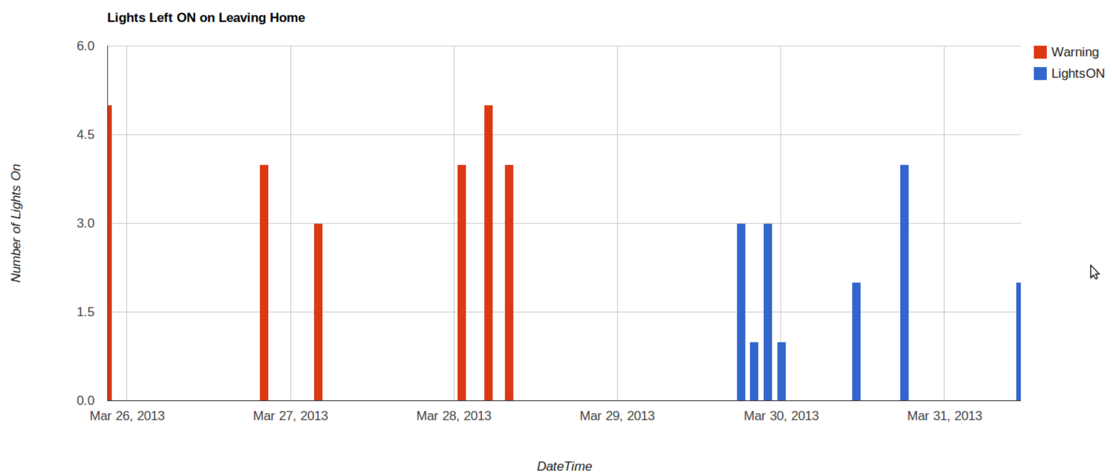


Figure B.3: Example of lights on visualization

B.4 Leaves visualization

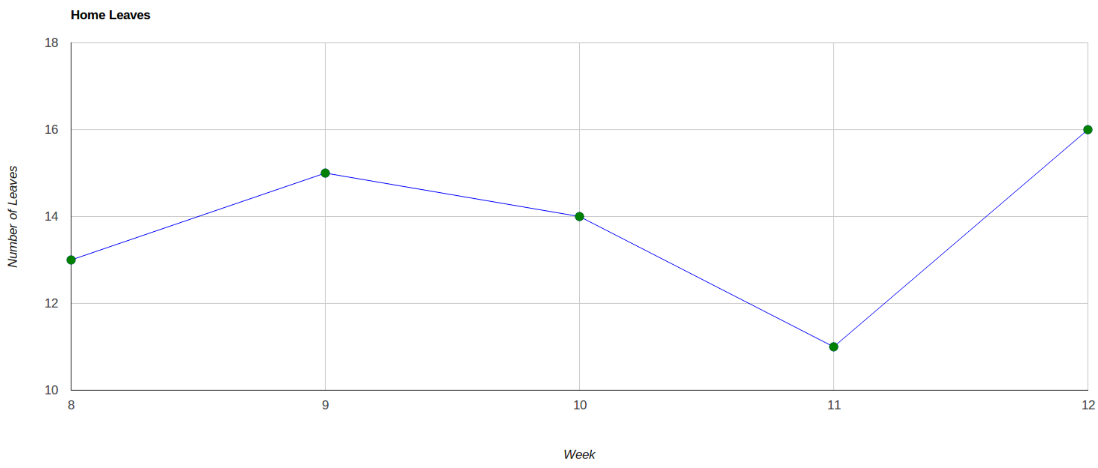


Figure B.4: Example of leaving home visualization

B.5 Frequent Patterns visualization

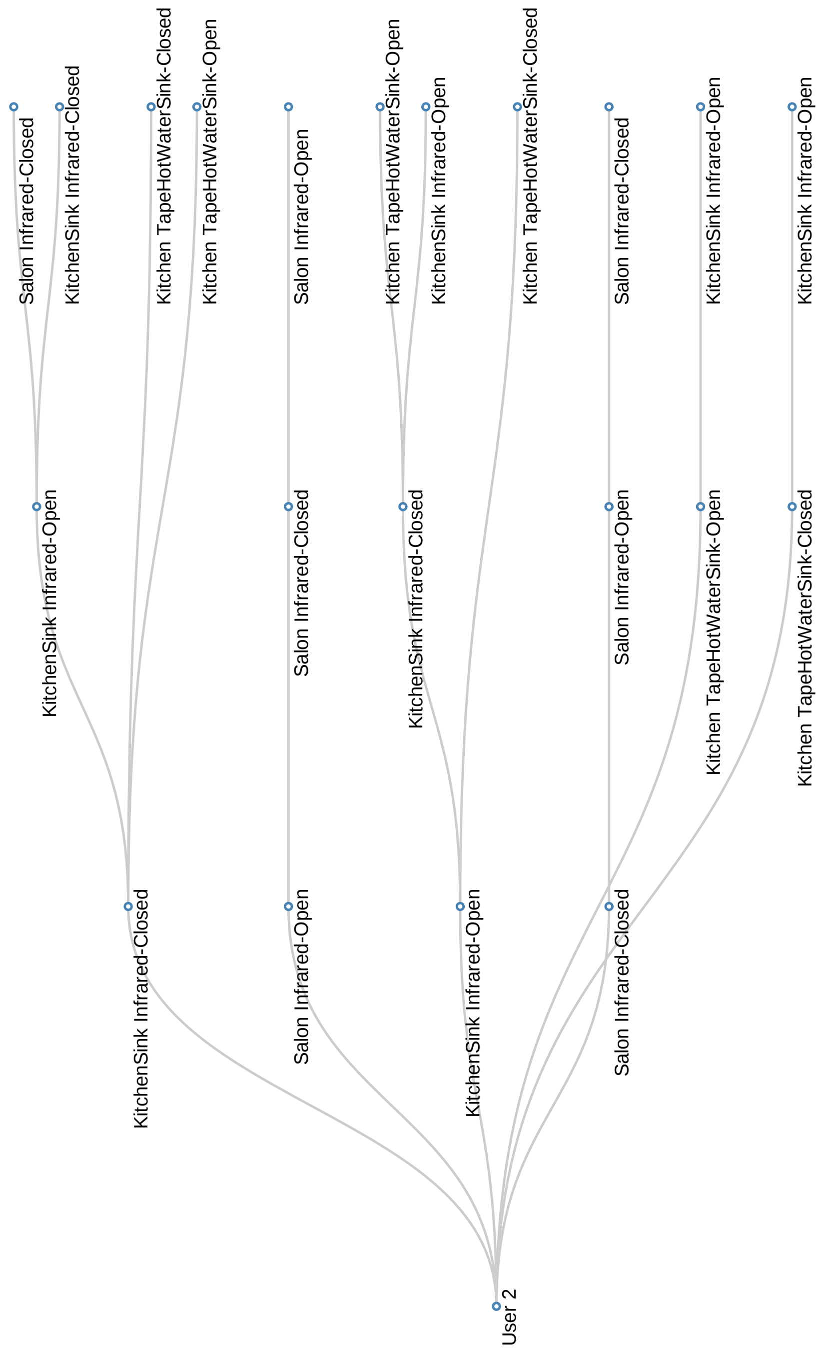


Figure B.5: Example of frequent sequential patterns visualization